Niveau(x) d'enseignement: 7 Numéro EDUQ : 8170451
Centre d'intérêt: 33

Bureau de la statistique du Québec. Service des méthodes statistiques.

La pensée formelle chez les étudiants de Collège I : objectif ou réalité? : rapport méthodologique réalisé pour le Ministère de l'éducation / Bureau de la statistique du Québec, Service des méthodes statistiques. -- [Québec] : Le Bureau, 1981.
87 f. : graph. ; 28 cm.

Bibliographies à la fin de la plupart des chap.

Cote Dewey: 370.152/B952p Cote LC: LB/1060

Descripteur(s): *Développement de l'étudiant *Développement intellectuel *Pensée formelle *Rendement scolaire *Analyse statistique *Collégial (Niveau)

Identificateur(s) de nom: *ERF (Test)

Le document expose la méthodologie statistique sur laquelle reposent les résultats présentés par Mirette Torkia-Lagacé dans le rapport ##La pensée formelle chez les étudiants de collège I : objectif ou réalité?## (1981). Cette étude, fruit d'une enquête menée auprès des étudiants de la région de Québec lors de leur inscription à l'automne 1979, a pour cadre méthodologique la théorie de Jean Piaget sur le développement intellectuel. Elle a pour objectif de découvrir, à l'aide d'un test, à quel niveau d'opération intellectuelle les étudiants fonctionnent et dans quelle mesure la capacité d'opérer au niveau de la logique formelle est associée à la réussite scolaire.

Après avoir décrit dans un premier chapitre les formules d'estimation statistique, le document présente cinq séries d'analyses: 1. Analyse du test ERF (Epreuve de raisonnement formel), 2. Analyse factorielle des correspondances appliquée aux résultats du test, 3. Analyse des caractéristiques personnelles par un modèle log-linéaire, 4. Analyse factorielle des correspondances entre le classement au test et les résultats scolaires, 5. Analyse des résultats scolaires selon le classement à la partie "énigmes" du test par des analyses de variance.

Chacune des cinq séries d'analyse est traitée en un chapitre qui présente les résultats à la suite de la description détaillée des données et de la méthode d'analyse. Pour une présentation d'ensemble des résultats de l'étude, le lecteur est renvoyé au rapport de Mirette Torka-Lagacé.

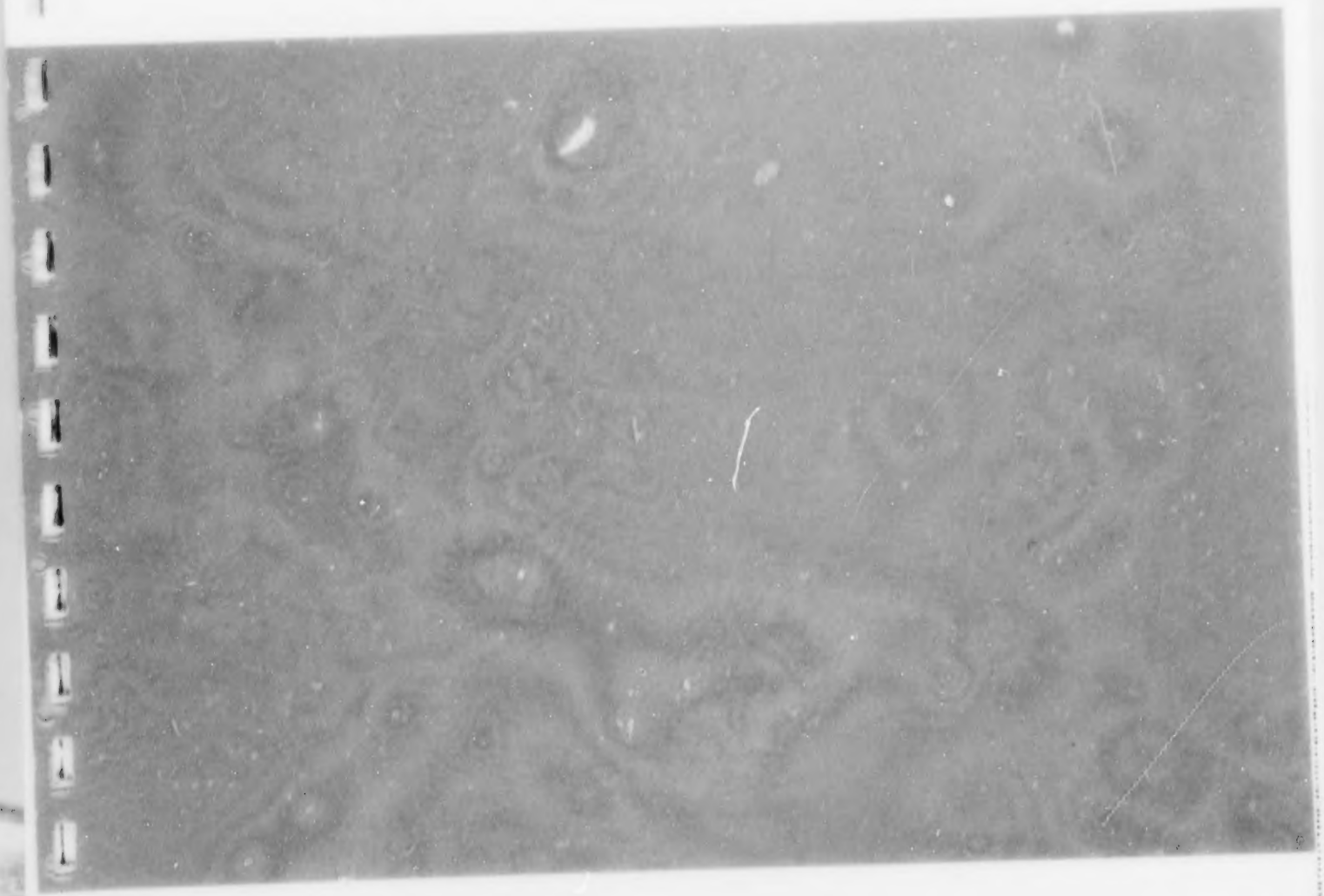
Type de document: r
Région(s) administrative(s): 063
Notice(s) connexe(s): 8170458
Auteur de l'analyse: ar

E D U Q 8170451

Bureau de
la statistique
du Québec

LA PENSÉE FORMELLE CHEZ LES ÉTUDIANTS DE COLLÈGE I - OBJECTIF OU RÉALITÉ ?

RAPPORT MÉTHODOLOGIQUE RÉALISÉ POUR
LE MINISTÈRE DE L'ÉDUCATION



E D U Q 8170451

LA PENSÉE FORMELLE CHEZ LES ÉTUDIANTS DE COLLÈGE I - OBJECTIF OU RÉALITÉ ?

RAPPORT MÉTHODOLOGIQUE RÉALISÉ POUR LE MINISTÈRE DE L'ÉDUCATION

BUREAU DE LA STATISTIQUE DU QUÉBEC
SERVICE DES MÉTHODES STATISTIQUES

JANVIER 1981

AVANT-PROPOS

Dans le cadre du projet " La pensée formelle chez les étudiants de collège I: objectif ou réalité ? " plusieurs problèmes de nature statistique se posaient avec ampleur et n'offraient pas de solution immédiate. Notre but fut d'élaborer les analyses statistiques et les analyses des données susceptibles d'apporter à la recherche l'éclairage objectif que l'on attendait de la masse des données recueillies. Le présent rapport présente donc la méthodologie statistique sur laquelle repose l'ensemble des résultats présentés par Mirette Torkia-Lagacé dans son rapport*. Il contient tous les renseignements nécessaires à celui qui voudrait refaire et comprendre la démarche de réflexion conduisant aux résultats statistiques obtenus. Il faudrait prendre garde d'y voir un exposé exhaustif des techniques statistiques utilisées (quantité de volumes existent déjà) pas plus qu'une présentation complète des résultats. Le lecteur intéressé pourra consulter le rapport de M. Lagacé à ce dernier sujet.

Plusieurs personnes ont contribué à ce rapport. Au Bureau de la statistique du Québec, Louise Bourque a mis au point les formules d'estimation, Paul Simard a effectué l'analyse du test et Claude Boivin les analyses des correspondances sur le test et les résultats scolaires. Le modèle log-linéaire et les analyses de la variance sont l'oeuvre de Marc Lavoie, consultant pour le Cegep de Limoilou. Il faut souligner l'apport précieux que Mirette Lagacé a apporté tout au long des travaux statistiques. Quant à la dactylographie du texte final, elle a été effectuée par Nicole Carlos que nous remercions.

---

* Mirette Torkia-Lagacé. La pensée formelle chez les étudiants de collège I: objectif ou réalité ? Cegep de Limoilou, 1981.

TABLE DES MATIÈRES

AVANT-PROPOS

TABLE DES MATIÈRES

CHAPITRE		Page
I	INTRODUCTION	1
II	FORMULES D'ESTIMATION	3
III	ANALYSE DE L'ÉPREUVE E.R.F.	8
IV	ANALYSE FACTORIELLE DES CORRESPONDANCES APPLIQUÉE AUX RÉSULTATS AU TEST	34
V	ANALYSE DES CARACTÉRISTIQUES PERSONNELLES PAR UN MODÈLE LOG-LINÉAIRE	52
VI	ANALYSE FACTORIELLE DES CORRESPONDANCES ENTRE LE CLASSEMENT AU TEST ET LES RÉSULTATS SCOLAIRES	66
VII	ANALYSE DES RÉSULTATS SCOLAIRES SELON LE CLASSEMENT À LA PARTIE ÉNIGMES DU TEST PAR DES ANALYSES DE LA VARIANCE ...	80

I - INTRODUCTION

On constate depuis plusieurs années des difficultés d'apprentissage chez les étudiants du niveau collégial, notamment dans l'apprentissage de la physique. On ne sait à quoi attribuer les échecs d'un grand nombre d'étudiants lors de chaque session. Les recherches menées aux Etats-Unis sur le sujet ont permis d'envisager certaines hypothèses de travail, hypothèses qu'ont quelque peu précisé des recherches menées au Québec ces dernières années.

Le projet de Mirette Lagacé tente de répondre aux questions suivantes*:

" - A quel niveau d'opération intellectuelle les étudiants admis en collège I fonctionnent-ils ?

- Dans quelle mesure la capacité d'opérer au niveau de la logique formelle est-elle associée à la réussite scolaire en général, et, en particulier, à la réussite scolaire dans des matières scientifiques telles que la physique et les mathématiques ? "

Pour atteindre ses objectifs, M. Lagacé a choisi de construire un instrument de mesure à passation collective (test papier-crayon) avec comme cadre méthodologique général la théorie de Jean Piaget sur le développement intellectuel et comme cadre spécifique le stade de la pensée opératoire formelle. Le test** est composé de trois parties: la première comporte un

---

* Mirette Torkia-Lagacé. La pensée formelle chez les étudiants de collège I: objectif ou réalité ? Cegep de Limoilou, 1981, p. 3.

** Pour une description complète, voir M. Lagacé, op. cit., chap. 2.

ensemble de tâches caractéristiques de la structure combinatoire; ces tâches sont subdivisées selon quatre stades (concret I (CI), concret II (CII), formel I (FI), formel II (FII)). La seconde partie est divisée en deux sous-parties: les proportions qui comportent des tâches des quatre stades (CI, CII, FI, FII) et les pourcentages dont les tâches sont caractérisées " concrètes " ou " formelles ". Enfin la troisième partie, appelée " Enigmes ", porte sur les critères généraux de la pensée formelle et devrait mesurer les raisonnements que les étudiants sont appelés à faire; cette partie comporte des tâches " concrètes " et " formelles " qui sont subdivisées par la suite (CI, CII, FI, FII). Le test dont il est question est appelé " Epreuve de raisonnement formel " (en abrégé " épreuve E.R.F. "). L'épreuve a été soumise à l'ensemble des étudiants admis en collégial I dans la région 03. De plus, on a adjoint aux résultats du test les résultats de chaque étudiant à un ensemble de cours des niveaux secondaire et collégial ainsi qu'un ensemble de caractéristiques personnelles décrivant cet étudiant.

Les questions faisant appel aux méthodes de la statistique et de l'analyse des données étaient nombreuses. Chacun des problèmes soulevés a reçu un traitement approfondi et fait l'objet d'un rapport. Le présent document rassemble ces rapports selon un ordre qui respecte au mieux la chronologie des analyses. Cependant, chaque chapitre est complet par lui-même et peut être consulté indépendamment des autres. Signalons rapidement les sujets étudiés: estimation (chap. 2), fidélité et validité de l'épreuve E.R.F. (chap. 3), analyse des résultats à l'épreuve E.R.F. (chap. 4 et 5) et lien entre les résultats au test et les résultats scolaires (chap. 6 et 7).

II - FORMULES D'ESTIMATION

L'étude sur la pensée formelle chez les étudiants de collège I consiste à évaluer le niveau de maturité intellectuel des étudiants à la rentrée au collégial, avec pour cadre de référence, la théorie de Piaget.

Un questionnaire a été distribué à tous les étudiants de la région de Québec*, en première année de collégial, lors de leur inscription à l'automne 1980. On désire obtenir des compilations de résultats selon divers sous-groupes de la population cible, et on s'intéresse surtout à la proportion d'étudiants possédant certaines caractéristiques.

Nous exposerons dans ce qui suit, les formules pour les estimateurs de proportion à divers niveaux, accompagnés de leurs estimateurs de la variance et de leurs intervalles de confiance.

---

* A l'exception de ceux inscrits dans deux (2) institutions non répondantes.

Soit une caractéristique X définissant un certain individu. Posons donc, pour chaque étudiant ayant répondu au questionnaire:

$$X_{srtji} = \begin{cases} 1 & \text{si cet individu i, étudiant à l'institution j} \\ & \text{dans la catégorie (champ d'étude)t du secteur r} \\ & \text{et de sexe s,possède la caractéristique X,} \\ 0 & \text{sinon,} \end{cases}$$

avec

$i = 1, \ldots, n_{srtj}$ (le nombre total d'étudiants ayant répondu au questionnaire, inscrits à l'institution j, dans la catégorie t du secteur r et de sexe s)

$j = 1, \ldots, 10$

$t = 1, \ldots, n_r$ (le nombre total de catégories qui ont été définies pour l'étude)

$r = 1, 2$

$s = 1, 2$

Spécifions que l'ensemble des étudiants inscrits à l'institution j dans la catégorie t du secteur r et de sexe s forme une strate. Au départ, il avait été prévu de faire un recensement parmi les N_{srtj} individus de chaque strate (c'est-à-dire de chaque institution j dans chaque catégorie t, de chaque secteur r et de chaque sexe s). En fait, de ces N_{srtj}, n_{srtj} ont répondu au questionnaire. Nous allons considérer que ces n_{srtj} unités sont réparties aléatoirement parmi la population, et nous établirons les formules d'estimation comme si on avait utilisé un échantillonnage simple au hasard dans chaque strate.

On désire savoir quelle est la proportion d'étudiants possédant la caractéristique X, selon divers niveaux. Nous présenterons donc les estimateurs de cette proportion accompagnés de leurs estimateurs de la variance et de leurs intervalles de confiance au niveau 0,95.

a) Au niveau de chaque catégorie

La proportion d'étudiants possédant la caractéristique X, pour chaque catégorie t du secteur r et de sexe s, s'établit comme suit:

$$\hat{p}_{srt} = \frac{\left[\sum_{j=1}^{10} \sum_{i=1}^{n_{srtj}} W_{srtj} \; X_{srtji}\right]}{N_{srt.}}$$

où

$$W_{srtj} = \frac{N_{srtj}}{n_{srtj}} \qquad \text{(poids dû à la non-réponse)}$$

et où

$$N_{srt.} = \sum_{j=1}^{10} N_{srtj} \; .$$

La variance associée à cet estimateur s'estime à l'aide de:

$$\widehat{var}(\hat{p}_{srt}) = \frac{\left[\sum_{j=1}^{10} W_{srtj} \; (W_{srtj} - 1) \; \frac{(n_{srtj})^2}{(n_{srtj} - 1)} \cdot \hat{p}_{srtj}(1 - \hat{p}_{srtj})\right]}{(N_{srt.})^2} \; .$$

où

$$\hat{p}_{srtj} = \frac{\sum_{i=1}^{n_{srtj}} W_{srtj} \; X_{srtji}}{N_{srtj}}$$

L'intervalle de confiance au niveau 0,95 pour $\hat{P}_{srt}$ se définit comme suit:

$$P\left(\left| P_{srt} - \hat{P}_{srt} \right| \leq e \right) = 0,95$$

avec

$$e = 1,96 \cdot \sqrt{\hat{var}(\hat{p}_{srt})}$$

P_{srt} étant la valeur réelle de la proportion dans la population.

b) Au niveau d'un certain ensemble de strates

La proportion d'étudiants possédant la caractéristique X, pour un certain ensemble L de strates h, h = 1, ..., ℓ, s'obtient ainsi:

$$\hat{P}_L = \sum_{h=1}^{\ell} \left(\frac{N_h}{N} \right) \hat{p}_h$$

où

$\hat{P}_h$ est la proportion d'étudiants possédant la caractéristique X dans la strate h

N_h est le nombre total d'étudiants appartenant à la strate h

et

$$N = \sum_{h=1}^{\ell} N_h .$$

On peut aussi obtenir l'estimateur de la variance de $\hat{P}_L$ de la façon suivante:

$$\widehat{\text{var}}(\hat{p}_L) = \sum_{h=1}^{\ell} \left[\left(\frac{N_h}{N} \right)^2 \widehat{\text{var}}(\hat{p}_h) \right] .$$

De là, on peut établir quel sera l'intervalle de confiance au niveau 0,95 pour $\hat{p}_L$:

$$P (| p_L - \hat{p}_L | \leq e) = 0,95$$

avec

$$e = 1,96 \cdot \sqrt{\widehat{\text{var}}(\hat{p}_L)}$$

p_L étant la valeur réelle de la proportion dans la population.

Les différents ensembles L de strates qui peuvent être considérées sont par exemple:

- l'ensemble des catégories d'un secteur par sexe,
- l'ensemble des secteurs par sexe,
- l'ensemble des secteurs pour les deux sexes.

III - ANALYSE DE L'ÉPREUVE E.R.F.

L'analyse de l'épreuve E.R.F. que nous avons effectuée, avait pour but de vérifier d'une part, que l'ensemble des questions du test présente les caractéristiques désirées en terme de fidélité des résultats observés et, d'autre part, si ces résultats sont en accord avec la théorie du développement intellectuel sous-jacente à la conception du test. Notre analyse est confirmatoire puisqu'elle n'avait pas pour but d'identifier des questions qui pourraient servir à construire un test mais plutôt d'évaluer la performance d'un instrument de mesure déjà constitué et dont la conception a fait l'objet d'une démarche exploratoire sérieuse.

Nous nous sommes basés sur le coefficient de consistance interne de Kuder et Richardson pour effectuer l'analyse de la fidélité du test. Quant à l'analyse de la validité du test, elle repose sur l'analyse hiérarchique en ce qui a trait à l'unidimensionnalité des résultats de chaque partie du test et sur l'analyse en composantes principales en ce qui a trait à l'interdépendance entre les parties du test.

1. L'ANALYSE DE LA FIDÉLITÉ DU TEST

L'analyse de la fidélité a pour but de déterminer l'importance des erreurs de mesure aléatoires dans les résultats observés ou encore de déterminer si les résultats auraient été à peu près semblables si on avait fait passer le test à un autre moment.

Les techniques usuelles d'analyse de fidélité reposent sur la théorie classique de la mesure et des tests en psychologie. Nous exposerons donc quelques éléments de cette théorie avant de donner les résultats de l'analyse de l'épreuve E.R.F. Cette présentation théorique ne se veut pas exhaustive et le lecteur trouvera dans Nunnaly [9, chap. 6 et 7] un exposé plus complet sur le sujet.

1.1 La théorie classique des tests et de la mesure

La théorie classique des tests et de la mesure est basée sur l'équation

$$x = t + e$$

où x représente la mesure du caractère étudié, t représente la valeur réelle de ce caractère et e est un terme d'erreur aléatoire qu'on suppose de moyenne nulle et indépendant du caractère, c'est-à-dire que

$$\text{et} \quad \begin{aligned} E(e) &= 0 \\ \text{cov}(t,e) &= 0 \end{aligned}$$

A partir de ces trois équations, on peut déduire que

$$\text{et} \quad \begin{aligned} E(x) &= E(t) \\ \text{Var}(x) &= \text{Var}(t) + \text{Var}(e) \end{aligned}$$

On voit donc, que selon ce modèle, plus la variance du terme d'erreur sera petite par rapport à celle du caractère observé, plus la mesure x sera fidèle pour le caractère t, c'est-à-dire qu'elle prendra à peu près toujours la même valeur sur le même individu. On définit donc du point de vue statistique, la fidélité de la mesure x pour le caractère t par

$$R_{xt} = 1 - (\text{Var}(e) / \text{Var}(x))$$

ou d'une façon équivalente par

$$R_{xt} = \text{Var}(t) / \text{Var}(e)$$

Un test est un instrument de mesure composé d'un ensemble de questions. Chacune de ces questions peut être représentée par une variable prenant la valeur 0 si la question n'est pas réussie et prenant la valeur 1 dans le cas contraire. Ainsi, une mesure obtenue à l'aide d'un test composé de k questions est une mesure déterminée par les valeurs que prennent un ensemble de variables binaires $\{x_1, \ldots, x_k\}$. Nous bornerons notre étude

aux cas où cette mesure est définie par $x = \sum_{i=1}^{k} x_i$. Dans ce cas particulier, on peut démontrer que la corrélation entre la variable x et le caractère t est donné par

$$r_{xt} = (k / (k - 1)) (1 - (\sum_{i=1}^{k} p_i q_i) / \sigma_x^2)$$

où p_i est la probabilité de réussite de l'item i et $q_i = 1 - p_i$.

Le coefficient r_{xt} définit par l'équation précédente est appelé coefficient alpha de Kuder-Richardson ou d'une façon plus succincte $\alpha_{KR - 20}$. On peut démontrer (voir [5]) que ce coefficient fournit une borne inférieure pour le coefficient de fidélité d'un test, c'est-à-dire que $R_{xt} \geq r_{xt}$.

L'estimation du coefficient $\alpha_{KR - 20}$ étant relativement simple et ne nécessitant qu'une seule passation du test, ce coefficient peut donc être utilisé facilement pour analyser la fidélité d'un test. C'est à partir de telles considérations que nous avons choisi de baser l'analyse de la fidélité du test E.R.F. sur le coefficient $\alpha_{KR - 20}$. L'inconvénient d'une telle démarche est qu'il n'existe pas de critère statistique pour juger si une valeur observée d'un coefficient $\alpha_{KR - 20}$ indique une grande fidélité. Cependant, l'usage courant est de considérer que la fidélité d'un test est très bonne lorsque le coefficient $\alpha_{KR - 20}$ est supérieur à 0,8. Notons que cette norme est souvent considérée comme étant très sévère.

1.2 Les résultats de l'analyse de la fidélité

L'épreuve E.R.F. est constituée de quatre parties: la combinatoire, la proportionnalité, les pourcentages et les énigmes. Chacune de ces parties mesure un aspect particulier du développement intellectuel et

l'étudiant est caractérisé par son profil de réponses à chacune des parties du test. Ces considérations nous ont amené à effectuer l'analyse de la fidélité à deux niveaux; au niveau de l'ensemble du test pour juger de sa capacité de mesurer globalement le développement intellectuel, et au niveau de chaque partie du test pour évaluer la qualité de la mesure de chaque aspect du développement intellectuel. Les calculs ont été effectués à l'aide de l'option RELIABILITY du logiciel SPSS. On a obtenu les résultats suivants:

1) Pour l'ensemble du test

$$\alpha_{KR - 20} = 0,921$$

2) Pour chaque partie du test

combinatoire	$\alpha_{KR - 20} = 0,770$
proportionnalité	$\alpha_{KR - 20} = 0,885$
pourcentages	$\alpha_{KR - 20} = 0,817$
énigmes	$\alpha_{KR - 20} = 0,807$

Ces résultats indiquent donc une bonne performance du test et nous permettent d'entreprendre sans problème des analyses statistiques basées sur les résultats du test. Notons cependant que le coefficient $\alpha_{KR - 20}$ de la partie combinatoire est légèrement inférieur à 0,8 la norme générale citée. On devra donc éviter d'effectuer des analyses trop élaborées en n'utilisant que les résultats à cette partie du test.

2. L'ANALYSE HIÉRARCHIQUE DE L'ÉPREUVE E.R.F.

L'analyse hiérarchique de l'épreuve E.R.F. a pour but de

vérifier si les résultats du test sont en accord avec la théorie du développement intellectuel quant à l'ordre d'accession aux différents stades de la pensée formelle. Cette analyse repose sur les techniques d'analyses des scalogrammes introduites par Louis Guttman.

L'analyse hiérarchique de l'épreuve E.R.F. est très liée à l'interprétation des différents profils de réponses et à la grille de correction utilisée pour classer les étudiants aux différents stades du développement de l'intelligence. Nous discuterons donc, après avoir exposé les principes de base de l'analyse hiérarchique, de cette relation entre la conception du test et la théorie du développement intellectuel pour bien situer l'analyse que nous avons effectuée. Nous présenterons, enfin, les résultats obtenus.

2.1 Les principes de base de l'analyse hiérarchique

L'analyse hiérarchique permet de déterminer si les questions d'un test ne mesurent qu'une seule grandeur ou, autrement dit, si le phénomène observé à l'aide de ces questions est unidimensionnel.

Considérons un test composé de k questions. On appellera score d'un individu au test, le nombre de questions que l'individu aura réussi. On dira alors que les questions du test forment une échelle au sens de Guttman, ou plus simplement échelle, si on peut reconstruire parfaitement le profil des réponses d'un individu en ne connaissant que son score au test. Considérons, par exemple, le test composé des trois questions suivantes:

1) Avez-vous plus de 20 ans ?

2) Avez-vous plus de 30 ans ?

3) Avez-vous plus de 40 ans ?

Dans ce cas, on dira qu'un individu a réussi une question lorsqu'il aura

répondu oui à celle-ci. Il est clair alors que nos trois questions forment une échelle au sens de Guttman. Par exemple, un individu qui a obtenu un score de 2 a nécessairement répondu oui aux deux premières questions et non à la troisième.

Le concept d'échelle que nous venons de définir ne fait intervenir aucune notion d'ajustement ou d'écart entre les profils de réponses reconstruits à partir des scores et ceux observés. Il s'agit d'un concept complètement déterministe et il est alors simple de vérifier si des questions $q_1, \ldots, q_k$ forment une échelle. Il suffit pour cela de vérifier s'il existe une permutation $q_{i_1}, \ldots, q_{i_k}$ de ces questions qui satisfait la condition suivante: pour tout $\ell = 1, \ldots, k$, si la question q_{i_k} est réussie, la question $q_{i_{\ell-1}}$ l'est aussi et si la question q_{i_ℓ} n'est pas réussie alors la question $q_{i_{\ell+1}}$ n'est pas réussie. Supposons, par exemple, qu'on a fait passer un questionnaire comportant quatre questions et que l'on a observé le profil de réponses du tableau 2.1.

Tableau 2.1: Profil des réponses

	Question			
Indidivu	q_1	q_2	q_3	q_4
1	1	0	1	0
2	1	1	1	0
3	0	0	1	0
4	1	0	1	1
5	0	0	0	0
6	0	0	1	0

Pour vérifier qu'il existe une permutation des questions vérifiant la condition qu'on vient d'énoncer, on construit un nouveau tableau en interchangeant les lignes du tableau initial de façon à ordonner les individus selon leur score,et en interchangeant les colonnes du tableau de façon à ce que les questions soient ordonnées selon leur degré de difficulté observé. Un tel tableau est appelé scalogramme des profils de réponses. Le tableau 2.2 nous donne le scalogramme qu'on peut construire à partir du tableau 2.1.

Tableau 2.2: Scalogramme des profils de réponses

	Question			
Individu	q_3	q_1	q_4	q_2
5	0	0	0	0
3	1	0	0	0
1	1	1	0	0
6	1	1	0	0
4	1	1	1	0
2	1	1	1	1

Une simple observation de ce scalogramme permet de vérifier que la permutation q_3, q_1, q_4, q_2 vérifie la condition énoncée plus haut et donc que ces quatre questions forment une échelle. Il suffit pour cela de constater qu'il n'y a aucune occurrence du patron 01 dans les profils de réponses présentés dans le scalogramme.

Le modèle déterministe des échelles de Guttman n'est, à toute fin pratique, jamais observé. Même lorsqu'un questionnaire mesure un caractère unidimensionnel, les erreurs de mesure amènent des profils de

réponses qui contredisent l'unidimensionnalité. Donc, pour effectuer une analyse hiérarchique réaliste, on doit considérer la possibilité de faibles écarts entre les profils de réponses observés et ceux conformes au modèle de Guttman. Notre problème est alors de définir une mesure de cet écart entre nos observations et le modèle théorique et de déterminer une norme permettant de juger la valeur de ces écarts.

Guttman [6] a proposé de mesurer l'importance de l'écart entre une échelle théorique et les profils de réponses de la façon suivante. Considérons, pour fixer les idées, un test composé des questions $q_1, \ldots, q_k$ ordonnées par ordre croissant de difficulté. Pour un profil de réponses dont le score est ℓ $(0 \leq \ell \leq k)$, on considèrera qu'il y a erreur si $q_i = 0$ pour $i \leq \ell$ ou $q_i = \ell$ pour $i > \ell$. Ainsi, pour $k = 5$, le profil 11010 contient deux erreurs; en effet, puisque le score est 3, on aurait dû avoir un 1 à la troisième question et un 0 à la quatrième. Soit E le nombre total d'erreurs observées dans l'ensemble des profils de réponses de n individus pour un test de k questions, alors le coefficient proposé par Guttman pour mesurer l'ajustement des profils au modèle est donné par

$$\text{rep} = 1 - \frac{E}{n\,k} .$$

Cette mesure est appelée coefficient de reproductibilité.

La valeur maximum que peut prendre le coefficient de reproductibilité est 1 et sa valeur minimum, comme d'ailleurs sa distribution, dépend du degré de difficulté des questions du test. Ne pouvant exprimer d'une façon générale la distribution du coefficient de reproductibilité, on doit donc se fier à l'usage pour juger de sa valeur et accepter arbitrairement la norme de 0,9 comme seuil à partir duquel on considère qu'un ensemble de questions s'ajuste assez bien au modèle des échelles de Guttman.

La variabilité de la distribution du coefficient de reproductibilité en fonction du degré de difficulté des questions nous empêche de

l'utiliser pour comparer différents ensembles de questions quant à l'ajustement du modèle des échelles de Guttman. Plusieurs auteurs ont donc proposé d'autres mesures d'ajustement au modèle hiérarchique. On trouvera en [10] une présentation assez complète de ces différents coefficients proposés. Quant à nous, la disponibilité d'un programme permettant de calculer facilement le coefficient de scalabilité, nous a incité à choisir ce coefficient comme mesure standard de l'ajustement au modèle hiérarchique.

On peut démontrer que pour un ensemble de questions dont les proportions de réussite sont données par $p_1, \ldots, p_k$, le coefficient de reproductibilité ne peut être inférieur à

$$rep^* = \left(\sum_{i=1}^{k} \max (p_i , 1 - p_i) \right) / n\, k$$

où max $(p_i, 1 - p_i)$ est le maximum entre p_i et $1 - p_i$. On définit alors le coefficient de scalabilité par

$$scal = (rep - rep^*) / (1 - rep^*).$$

Le coefficient de scalabilité peut varier entre 0 et 1 quel que soit les pourcentages de réussite aux questions et prend la valeur 1 lorsque l'ajustement est parfait. On considère ordinairement qu'un coefficient de scalabilité supérieur à 0,7 indique un très bon ajustement au modèle des échelles de Guttman.

2.2 L'application de l'analyse hiérarchique au test E.R.F.

L'épreuve E.R.F. a pour but d'identifier les raisonnements qu'un étudiant maîtrise et qui caractérisent la pensée formelle. A cette fin, lors de la conception du questionnaire, on a conçu, pour chaque stade du développement intellectuel, un ensemble de questions permettant de déterminer si l'étudiant maîtrise les concepts caractéristiques du stade en question. Ainsi, au niveau de la combinatoire, la maîtrise du calcul des com-

binaisons de deux éléments parmi n caractérise le stade concret II. On a donc posé à l'étudiant des questions dans lesquelles n'intervient que le calcul des combinaisons de deux éléments parmi n. Ainsi, chaque question du test est identifiée à un stade du développement intellectuel. Lors de la correction de l'épreuve et de l'interprétation des profils de réponses, on a défini des critères pour déterminer si le profil des réponses de l'étudiant nous assure qu'il maîtrise les concepts caractéristiques de chaque stade. On trouvera dans le texte de Mirette Lagacé [7] la description de ces critères. Il est important de noter que ceux-ci n'ont pas été déterminés à l'aide de techniques statistiques mais plutôt à l'aide de la théorie du développement intellectuel qui est à la base de l'ensemble de la recherche.

L'analyse hiérarchique a été utilisée pour déterminer si dans chaque partie du test, la maîtrise des différents stades constitue une échelle. Les résultats de cette analyse auraient pu être différents si les seuils de maîtrise de chacun des stades avaient été définis autrement. On aurait pu obtenir ainsi des coefficients de scalabilité et de reproductibilité artificiellement élevés. Cependant, il serait difficile de défendre des seuils de maîtrise qui n'auraient été définis qu'en fonction de la valeur de ces deux coefficients.

Comme on l'a indiqué précédemment, l'épreuve E.R.F. est divisée en quatre parties, chacune de ces parties mesurant un aspect particulier du développement de l'intelligence. Dans les parties combinatoire et proportionnalité, les questions caractérisaient quatre stades du développement qui étaient désignés par concret I, concret II, formel I et formel II. Pour la partie des pourcentages, on ne considère que deux stades: le stade concret et le stade formel. Enfin pour la partie des énigmes, les questions caractérisaient les stades concret I, concret II, formel I et formel II. Selon la théorie piagétienne, la maîtrise des tâches caractéristiques de chacun des stades se fait de façon hiérarchique, c'est-à-dire que, par exemple, il est

théoriquement impossible de maîtriser le stade formel I à la combinatoire si on ne maîtrise par les stades concret I et concret II de cet aspect du développement intellectuel. On voit donc que, pour chaque partie du test, la maîtrise des tâches caractéristiques des stades du développement devrait être modélisée par une échelle de Guttman.

Pour effectuer l'analyse hiérarchique de l'épreuve E.R.F., on a construit, à partir des profils de réponses observés, quatre ensembles de variables binaires. Le premier ensemble se rapportait à la partie combinatoire et comprenait les variables C1, C2, C3, C4 associées respectivement aux stades concret I, concret II, formel I et formel II. Lorsqu'un étudiant avait satisfait les critères de maîtrise d'un stade du schème de la combinatoire, la variable qui lui était associée prenait la valeur 1. D'une façon semblable, on a construit pour la partie proportionnalité l'ensemble de variables {P1, P2, P3, P4} associées respectivement aux stades concret I, concret II, formel I et formel II. Pour la partie des pourcentages, on a associé au stade concret la variable P01 et, au stade formel, la variable P02. Enfin, pour la partie des énigmes, puisque les étudiants étaient au moins classés concret I et que la distinction entre les stades formel I et formel II ne reposait que sur le nombre de questions du stade formel réussies, on n'a construit que les questions E1 et E2 représentant respectivement le stade concret II et l'ensemble du stade formel. L'analyse hiérarchique de l'épreuve E.R.F. a alors consisté à analyser l'ajustement de ces quatre ensembles de questions au modèle des échelles de Guttman.

Nous avons utilisé l'option GUTTMAN SCALE du logiciel SPSS pour effectuer notre analyse. Celle-ci étant basée sur environ 6 000 profils de réponses, on comprendra facilement qu'il nous est impossible de présenter les scalogrammes des réponses pour chaque partie du test. De tels scalogrammes nécessiteraient chacun environ 6 000 lignes. Cependant, pour illustrer l'ajustement au modèle de Guttman, on présentera, pour chaque partie de l'épreuve, un tableau croisé dont les colonnes indiquent le plus haut

stade maîtrisé par l'étudiant et les lignes indiquent le plus haut stade maîtrisé pour lequel l'étudiant a aussi démontré la maîtrise des stades qui le précédent. Par exemple, dans un tel tableau, un étudiant qui aurait satisfait les critères des stades concret I, concret II et formel II mais n'aurait pas satisfait les critères du formel I ferait partie des effectifs représentés à la croisée de la colonne formel II et de la ligne concret II. Dans un tel tableau, on ne retrouve aucun élément en dehors de la diagonale lorsque l'ajustement aux échelles de Guttman est parfait. Notons enfin que lorsqu'un étudiant ne satisfaisait pas les critères du premier stade d'une échelle, on le classait alors dans le stade pré-opératoire sauf pour la partie des énigmes où il était alors classé concret I.

2.3 Les résultats de l'analyse hiérarchique

Pour la partie combinatoire, on a observé les résultats présentés au tableau 2.3.

Tableau 2.3: Tableau croisé de la partie combinatoire

	Pré-opératoire	Concret I	Concret II	Formel I	Formel II
Pré-opératoire	11	0	15	5	0
Concret I	0	201	0	50	2
Concret II	0	0	1 694	0	65
Formel I	0	0	0	2 655	0
Formel II	0	0	0	0	1 180

On ne note que 137 cas dans la partie supérieure du tableau, ce qui nous donne lorsqu'on utilise la pondération du modèle échantillonnal[1] 2,3% de profils de réponses disconfirmatoires. Pour cette partie du test, on obtient un coefficient de reproductibilité de 0,994 et un coefficient de scalabilité de 0,96. On ne peut donc douter de la hiérarchie de ces stades.

Le tableau 2.4 représente le tableau croisé de profils de réponses de la partie proportionnalité.

Tableau 2.4: Tableau croisé de la partie proportionnalité

	Pré-opératoire	Concret I	Concret II	Formel I	Formel II
Pré-opératoire	39	0	28	9	13
Concret I	0	412	0	167	2
Concret II	0	0	1 255	0	62
Formel I	0	0	0	1 429	0
Formel II	0	0	0	0	2 443

On n'observe que 281 profils disconfirmatoires, ce qui nous donne un pourcentage de 4,8% après pondération. Le coefficient de reproductibilité est 0,976 et le coefficient de scalabilité est 0,888. Encore une

1 Voir [3].

fois, la hiérarchie des stades ne peut être mise en doute.

Pour la partie des pourcentages, on n'a que trois stades. Le tableau croisé est donné par le tableau 2.5.

Tableau 2.5: Tableau croisé de la partie des pourcentages

	Pré-opératoire	Concret I	Concret II
Pré-opératoire	360	0	32
Concret I	0	1 664	0
Concret II	0	0	3 783

Comme dans le cas précédent, l'ajustement au modèle de Guttman, on n'observe que 0,5% de profils disconfirmatoires, un coefficient de reproductibilité de 0,995 et un coefficient de scalabilité de 0,978.

Dans la partie des énigmes, la distinction entre les stades formel I et formel II ne reposait que sur le nombre de questions identifiées au stade formel que l'étudiant réussissait. Il est donc évident qu'il y avait une hiérarchie entre ces deux stades. De plus, un étudiant était considéré comme concret I même s'il ne réussissait aucune question. Pour l'analyse hiérarchique, on ne considère donc que les stades concret I, concret II et formel. Le tableau 2.6 est le tableau croisé des stades de cette partie du test.

Tableau 2.6: Tableau de la partie des énigmes

	Concret I	Concret II	Formel
Concret I	2 648	0	302
Concret II	0	933	0
Formel	0	0	1 913

On n'a observé que 5,2% de profils disconfirmatoires. Pour cette partie du test, le coefficient de reproductibilité est 0,95 et le coefficient de scalabilité est 0,8813.

En résumé, on constate que l'accession aux différents stades du développement intellectuel tel que nous l'avons observé à l'aide de l'épreuve E.R.F. se fait en conformité avec la théorie piagétienne. On peut en conclure, que le test et les critères de maîtrise des différents stades qui ont été utilisés permettent de mesurer correctement le développement intellectuel.

3. L'ANALYSE FACTORIELLE DES PROFILS DE RÉPONSES

L'application de l'analyse factorielle en composantes principales à l'analyse de l'épreuve E.R.F. avait pour but d'étudier les interdépendances entre les parties du test. Plus précisément, on a voulu, dans une première analyse, étudier jusqu'à quel point les profils de réponses observés s'accordait à la théorie piagétienne quant à l'indépendance entre la structure

de la pensée combinatoire et le schème de la proportionnalité. Notre but était donc de rechercher deux facteurs indépendants sous lesquels on pourrait regrouper d'une part les questions de la partie combinatoire et d'autre part celles de la partie proportionnalité et de la partie des pourcentages. De plus, dans une deuxième analyse, on a voulu étudier les relations entre la partie des énigmes et les autres parties du test.

L'utilisation de l'analyse en composantes principales pour des données dichotomiques est peu courante et doit faire l'objet d'attentions particulières. Nous justifierons donc son utilisation pour l'analyse des profils de réponses après avoir fait un bref exposé théorique sur les composantes principales. Nous présenterons par la suite les résultats des analyses que nous avons effectuées.

3.1 L'analyse factorielle en composantes principales

L'analyse en composantes principales est une technique d'analyse de données multidimensionnelles basée sur la recherche de variables synthétiques permettant une représentation optimale des observations. Cette technique peut être utilisée soit dans le but d'obtenir une représentation géométrique des données permettant de visualiser les relations entre les variables, soit pour rechercher un ensemble de facteurs communs orthogonaux expliquant la plus grande partie possible de la variabilité des observations. Dans le premier cas, on parle alors d'approche géométrique aux composantes principales et on trouvera dans le livre de F. Caillez et J.-P. Pages [4] une présentation de cette démarche. Quant à nous, nous avons utilisé l'analyse en composantes principales pour rechercher des facteurs orthogonaux, c'est-à-dire des variables indépendantes se basant ainsi sur la deuxième approche qu'on peut appeler " approche classique de l'analyse factorielle en composantes principales ". La discussion théorique qui suit est donc basée sur cette philosophie. Pour une présentation plus complète, on pourra consulter le livre de D.F. Morisson [8] ou de T.W. Anderson [2].

Soit $\underline{X} = (X_1, \ldots, X_p)$ un vecteur aléatoire avec $E(\underline{X}) = \underline{\mu}$ et $E(\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})' = \Sigma$ définie positive. Puisqu'on ne s'intéresse qu'aux variances et covariances, on suppose, sans perte de généralité, que $\underline{\mu} = 0$. On définit alors la première composante principale du vecteur $\underline{X}$ comme étant la combinaison linéaire $C_1 = \beta_1 \underline{X}$ des variables $X_1, \ldots, X_p$ telle que $\beta_1' \beta_1 = 1$ et que la variance de C_1 soit maximale. La deuxième composante principale est alors la combinaison linéaire $C_2 = \beta_2' \underline{X}$ telle que $\beta_2' \beta_2 = 1$, C_2 est indépendante de C_1 et que la variance de C_2 soit la plus grand possible. Enfin, d'une façon générale, pour $1 < \ell \leq p$, la ℓ^e composante principale du vecteur $\underline{X}$ est la combinaison linéaire $C_\ell = \beta_\ell' \underline{X}$, pour laquelle $\beta_\ell' \beta_\ell = 1$, $\beta_k' \beta_\ell = 0$ pour $1 \leq k < \ell$ et dont la variance est maximale. Les vecteurs β_ℓ, $\ell = 1, \ldots, p$ sont appelés les facteurs principaux de $\underline{X}$.

On peut montrer (voir [2], chap. 11) que la première composante principale est obtenue en prenant pour vecteur β_1, le vecteur propre unitaire de la matrice Σ correspondant à la plus grande valeur propre. D'une façon générale, pour $1 \leq \ell \leq p$, la ℓ^e composante principale est obtenue en prenant pour vecteur β_ℓ le vecteur propre unitaire de Σ correspondant à la ℓ^e plus grande valeur propre. De plus, on peut démontrer que la variance de ℓ^e composante principale est donnée par

$$\text{Var}\ (C_\ell) = \text{Var}\ (\beta_\ell' \underline{X}) = \lambda_\ell$$

et que

$$\sum_{\ell=1}^{p} \lambda_\ell = \sum_{\ell=1}^{p} \sigma^2_{X_\ell} = \text{tr}\ \Sigma$$

où tr Σ est la trace de la matrice Σ.

Notons enfin que, pour $j, \ell = 1, \ldots, p$, la covariance entre X_j et C_ℓ est donné par

$$\text{Cov } (X_j, C_\ell) = \text{Cov } (\varepsilon_j' \underline{X}, \beta_\ell' \underline{X})$$

ε_j est le vecteur dont les p composantes sont nulles sauf la j^e qui est égale à 1. On a donc

$$\text{Cov } (X_j, C_\ell) = \varepsilon_j' \Sigma \beta_\ell = \lambda \beta_{j\ell}$$

où $\beta_{j\ell}$ est la j^e composante du facteur β_ℓ. Si $\sigma_{X_\ell} = 1$, puisque var (C_ℓ) $= \lambda_\ell$, la corrélation entre la variable X_j et la composante principale C_ℓ est donné par

$$r_{X_j, C_\ell} = (\lambda)^{\frac{1}{2}} \beta_\ell .$$

Ce coefficient de corrélation est aussi appelé coefficient de saturation, ou simplement saturation, de la variable X_j sur le ℓ^e facteur.

Au niveau de l'échantillon, on estime les facteurs principaux et les composantes principales à partir de la matrice des données observées dans laquelle on a enlevé à chaque observation d'une variable la valeur de la moyenne de cette variable. On obtient une nouvelle matrice d'observations

$$X = \begin{pmatrix} x_1^1 & \cdots & x_i^1 & \cdots & x_n^1 \\ \vdots & & & & \\ x_1^j & \cdots & x_i^j & \cdots & x_n^j \\ \vdots & & & & \\ x_1^p & \cdots & x_i^p & \cdots & x_n^p \end{pmatrix}$$

où x_i^j est la valeur prise par la variable centrée X_j sur le i^e individu.

A partir de cette matrice, on peut estimer Σ par $S = (1 / (n - 1))\ X' X$. On effectue donc l'analyse en composantes principales sur cette matrice [1] et on obtient les vecteurs $\hat{\beta}_1, \ldots, \hat{\beta}_p$ qui estiment les facteurs principaux. Les composantes principales sont alors estimées par $X \hat{\beta}_\ell$, $\ell = 1, \ldots, p$. Les variances échantillonnages des composantes principales sont données par les vecteurs propres $\hat{\lambda}_\ell$, $\ell = 1, \ldots, p$, de la matrice S et la somme des variances échantillonnales des variables $X_1, \ldots, X_p$ est égale à la somme des valeurs propres de la matrice S. On a donc

$$\sum_{j=1}^{p} \widehat{\text{Var}}\ X_j = \sum_{\ell=1}^{p} \hat{\lambda}_\ell = \text{tr } S.$$

On dira alors que le pourcentage de variation expliqué par la ℓ^e composante principale (ou ℓ^e facteur principal) est donné par $\hat{\lambda}_\ell$ / tr S et dans la pratique on ne tient compte que des facteurs dont le pourcentage d'explication est suffisamment fort.

L'analyse en composantes principales d'une matrice d'observations nous permet donc d'identifier les combinaisons linéaires des variables observées qui expliquent une très grande partie de la variabilité des observations. Cette analyse nous fournit ainsi un ensemble de variables indépendantes au niveau de l'échantillon. Il est cependant rare que l'analyse en composantes principales nous fournisse des facteurs qui sont simples à interpréter, chaque variable étant habituellement reliée plus ou moins fortement à plusieurs facteurs. Afin d'obtenir une représentation plus simple des données, on peut effectuer une rotation des facteurs initiaux et obtenir ainsi un nouvel ensemble de facteurs orthogonaux et donc de variables indépendantes, dont les profils de saturation sont moins complexes.

---

1 En pratique, on factorise habituellement la matrice des coefficients de corrélation plutôt que la matrice S. Cela a pour but de rendre notre analyse des variables indépendante des unités de mesure.

En analyse factorielle, on fait souvent l'hypothèse que les variables observées sont distribuées normalement. Cette hypothèse permet alors d'autres développements théoriques dont la définition de tests sur les valeurs propres de la matrice de variances - covariances. Cependant, notons que la présentation de l'analyse en composantes principales que nous avons faite ne repose sur aucune hypothèse concernant la distribution des variables observées. D'autre part, puisque l'analyse repose sur les variances et covariances ou les corrélations entre les variables, elle doit être effectuée sur des variables mesurées sur une échelle proportionnelle ou d'intervalles. Tel que l'a démontré Anderberg [1, p. 29], les données binaires peuvent être considérées comme étant mesurées sur des échelles d'intervalles et donc on peut appliquer l'analyse factorielle en composantes principales aux profils de réponses de l'épreuve E.R.F. Cependant, puisque d'une façon évidente l'hypothèse de normalité n'est pas respectée, on ne peut effectuer les tests classiques sur les valeurs propres obtenues.

3.2 Les résultats de l'analyse

Comme on l'a indiqué précédemment, on a utilisé, dans un premier temps, l'analyse factorielle pour étudier les relations entre les parties du test mesurant d'une part le développement du schème de la proportionnalité[1] et d'autre part de la pensée combinatoire. On a donc appliqué la technique de l'analyse en composantes principales au tableau de données constitué des profils des réponses des étudiants aux trois premières parties du test. On a ainsi exclus de cette analyse les résultats à la partie des énigmes afin d'identifier plus facilement des facteurs indépendants reliés aux composantes combinatoire et proportionnalité de la pensée formelle. Le tableau de données analysé comprenait donc 5 878 profils de 50 variables binaires.

Les trois plus grandes valeurs propres obtenues par la factorisation de la matrice des corrélations sont 21,01 , 3,86 et 3,37 et nous indiquent que les trois premières composantes principales expliquent

---

1 Dans l'épreuve E.R.F., le développement du schème de la proportionnalité est mesuré par la partie proportionnalité et la partie des pourcentages.

respectivement 42,0%, 7,7% et 6,8% de la variabilité de nos observations. Nous avons choisi de ne pas considérer d'autres facteurs parce que d'une part, les trois premiers expliquent une partie suffisamment importante de la variabilité de nos observations et que d'autre part, comme on le verra plus loin, ils nous fournissent les résultats recherchés. Le tableau 3.1 indique les intervalles dans lesquels se situent les saturations des variables de chacune des parties sur les trois premiers facteurs principaux[1].

Tableau 3.1: Saturations des questions des trois premières parties du test sur les trois facteurs principaux

	Facteur 1	Facteur 2	Facteur 3
Questions de la partie combinatoire	(0,03 0,14)	(0,18 0,57)	(0,06 0,30)
Questions de la partie proportionnalité	(0,77 0,89)	(-0,09 0,17)	(-0,22 -0,03)
Questions de la partie des pourcentages	(0,67 0,69)	(-0,32 -0,42)	(0,49 0,57)

Le tableau 3.1 indique un comportement différent pour chaque partie du test et une cohérence à l'intérieur de chacune de ces parties. Le premier facteur peut être associé directement au schème de la proportionnalité puisque les questions de la partie proportionnalité et de la partie des pourcentages lui sont fortement correlées et que les saturations des questions de la partie combinatoire sont faibles. Le deuxième facteur oppose les questions de la partie combinatoire à celle des pourcentages tandis que le troisième facteur est relié principalement aux pourcentages. Bien que ces résultats montrent des différences importantes entre les parties

1 Par exemple, le tableau 3.1 indique qu'aucune des variables reliée à une question de la partie combinatoire a une saturation inférieure à 0,14 et supérieure à 0,03 sur le premier facteur.

du test, elles ne permettent pas d'identifier des facteurs reliés uniquement à la partie combinatoire ou aux parties du schème de la proportionnalité. Nous avons donc effectué une rotation, par la méthode du quartimax pour trouver de nouveaux facteurs. Le tableau 3.2 nous indique les intervalles dans lesquels se situent les saturations des questions sur chacun des nouveaux facteurs.

Tableau 3.2: Saturations des questions sur les trois facteurs obtenus par la rotation du quartimax

	Facteur 1	Facteur 2	Facteur 3
Questions de la partie combinatoire	(0,02 0,08)	(-0,04 0,04)	(0,20 0,65)
Questions de la partie proportionnalité	(0,78 0,90)	(-0,03 0,08)	(-0,09 0,19)
Questions de la partie des pourcentages	(0,50 0,54)	(0,76 0,79)	(-0,09 0,04)

L'analyse du tableau 3.2 nous permet d'identifier le premier facteur au schème de la proportionnalité tandis que le troisième peut être identifié à la pensée combinatoire. Enfin, le deuxième montre un élément particulier aux questions des pourcentages qui les distinguent des autres questions de la proportionnalité. Cette analyse nous permet donc d'identifier des facteurs reliés soit au schème de la proportionnalité soit à la pensée combinatoire illustrant ainsi une forme d'indépendance entre ces deux composantes de la pensée formelle.

Nous avons aussi utilisé l'analyse en composantes principales pour étudier les profils de réponses des étudiants à l'ensemble des questions du test. Cette démarche avait pour but d'identifier les liens entre la partie des énigmes et les autres parties du test. Les quatre plus grandes valeurs propres obtenues par la factorisation de la matrice des corrélations de l'ensemble des questions sont 23,79 , 6,28 , 3,76 et 2,25. Les pourcentages de variabilité expliqués par les quatre premières composantes principales sont respectivement 40,3%, 10,6%, 6,4% et 3,8%. Le tableau 3.3 indique les intervalles dans lesquels se situent les saturations des questions sur chacun des facteurs principaux retenus.

Tableau 3.3: Saturations des questions des différentes parties du test sur les quatre facteurs principaux

	Facteur 1	Facteur 2	Facteur 3	Facteur 4
Questions de la partie combinatoire	(0,02 0,14)	(-0,08 -0,03)	(0,18 0,64)	(-0,01 0,08)
Questions de la partie proportionnalité	(0,74 0,86)	(-0,29 -0,22)	(-0,17 0,11)	(-0,15 0,0)
Questions de la partie des pourcentages	(0,69 0,76)	(0,37 0,42)	(-0,13 -0,02)	(0,34 0,44)
Questions de la partie des énigmes	(0,58 0,62)	(0,63 0,65)	(0,01 0,1)	(-0,32 -0,29)

Le premier facteur nous montre directement le lien entre la partie des énigmes et le schème de la proportionnalité tandis que le deuxième facteur relie plus spécifiquement les questions des pourcentages aux énigmes. Le troisième facteur est identifié à la secteur combinatoire tandis que le quatrième oppose les énigmes aux pourcentages. On voit donc qu'il existe différentes relations entre les énigmes et les autres parties du test. Pour une discussion complète sur ce sujet, on pourra consulter le texte de Mirette Lagacé [7].

BIBLIOGRAPHIE

[1] Anderberg, M.R. Cluster Analysis for Applications. Academic Press, 1973.

[2] Anderson, T.W. An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, 1958.

[3] Bourque, L. La pensée formelle des étudiants de collège I: Formules d'estimation.

[4] Caillez, F., Pages, J.P. Introduction à l'analyse des données. SMASH, 1976.

[5] Cronbach, L.J. Coefficient Alpha and the Internal Structure of a Test. Psychometrika, vol. 16, no 3, 1951, pp. 297-334.

[6] Guttman, L.L. The Basis for Scalogram Analysis dans Stouffer, S.A. et al., Studies in Social Psychology in World War II. Princeton University Press, 1950, pp. 60-90.

[7] Torkia-Lagacé, M. La pensée formelle des étudiants de collège I; objectif ou réalité ? Cegep de Limoilou, 1981.

[8] Morisson, D.F. Multivariate Statistical Methods. Mc Graw-Hill, 1976.

[9] Nunnaly, J.C. Psychometric Theory. Mc Graw-Hill, 1967.

[10] Saltz, E., White, B.J. Measurement of Reproductibility. Psychological Bulletin, vol. 54, no 2, 1957, pp. 81-99.

IV - ANALYSE FACTORIELLE DES CORRESPONDANCES APPLIQUÉE AUX RÉSULTATS AU TEST

A l'automne 1979 les étudiants du collégial I de la région 03 ont été soumis à un test constitué de quatre parties (combinatoire, proportions, pourcentages, énigmes). Chaque individu est caractérisé par son classement à chacune des parties du test; il s'agit d'un classement en quatre stades (concret I, concret II, formel I, formel II). Nous avons utilisé l'analyse factorielle des correspondances pour étudier la relation qu'il peut y avoir entre les diverses variables de classement. En plus des variables de classement au test, nous possédons une description des individus par rapport à certaines caractéristiques personnelles (sexe, secteur, scolarité du père, etc.) que nous avons introduites dans l'analyse.

1. DESCRIPTION DÉTAILLÉE DES DONNÉES

Chacun des quelques 6 000 individus qui ont participé au test est caractérisé par son classement aux quatre parties. Le tableau 1 présente l'ensemble des variables de classement au test avec des étiquettes qui identifient chaque modalité. Dans les trois premières parties du test, un individu ne maîtrisant aucun des quatre stades était classé " préopératoire " par défaut. Dans la quatrième partie, l'individu était classé " concret I " par défaut s'il ne maîtrisait aucun des trois stades suivants. Par ailleurs, dans la partie " pourcentages " on classait les individus en fonction de deux stades seulement (concret, formel).

Tableau 1: Variables de classement au test

Identifieur	Partie	Stade
C0	Combinatoire	Pré-opératoire
C1	"	Concret I
C2	"	Concret II
C3	"	Formel I
C4	"	Formel II
PR0	Proportions	Pré-opératoire
PR1	"	Concret I
PR2	"	Concret II
PR3	"	Formel I
PR4	"	Formel II
P00	Pourcentages	Pré-opératoire
P0A	"	Concret
P0B	"	Formel
E1	Énigmes	Concret I
E2	"	Concret II
E3	"	Formel I
E4	"	Formel II

Au tableau 2, nous avons regroupé les caractéristiques personnelles qui seront examinées par rapport aux variables de classement au test.

Tableau 2: Caractéristiques personnelles

Identifieur	Question	Modalité
GAR	Sexe	Masculin
FIL		Féminin
GEN	Secteur	Général
PRF		Professionnel
SV1	Secondaire V terminé en	Non trouvé
SV2		1978-1979
SV3		1977-1978
SP0	Scolarité du père	Sans objet
SP1		0 à 3 ans d'études
SP2		4 à 6 " "
SP3		7 à 9 " "
SP4		10 à 12 " "
SP5		13 à 15 " "
SP6		16 à 19 " "
SP7		20 ans et plus
R1	Catégorie de programme	Sciences de la santé (010.00)*
R2		Sciences pures et appliquées (020.00)
R3		Sciences humaines sans mathématiques: Droit (031.00), Histoire (033.00), Philosophie (034.00), Communications (034.01), Sciences religieuses et théologie (037.00)
R5		Sciences humaines avec mathématiques: Géographie (032.00), Psychologie (035.00), Sciences sociales (038.00), Service social (038.01), etc.
R7		Sciences de l'administration (040.00)

* Numéro du programme au collégial.

Tableau 2: Caractéristiques personnelles (suite)

Identifieur	Question	Modalité
R8	Catégorie de programme (suite)	Arts (050.00), Lettres (060.00), Langues (067.01), Hors D.E.C. (070.00)
R9		Techniques d'hygiène dentaire (111.00), Techniques médicales (140.00), Techniques d'inhalothérapie et anesthésie (141.00), Techniques de sciences naturelles (145.00)
R10		Techniques de diététique (120.00)
R11		Techniques infirmières (180.00 et 180.21)
R12		Techniques de radiodiagnostic (142.00), Techniques de réadaptation (144.00), Techniques forestières (190.00)
R13		Techniques de chimie industrielle (210.00), Techniques de chimie biologique (210.03), Techniques de matières plastiques (211.00), Techniques de la mécanique (241.00), Electrotechnique (243.00), Technologie physique (244.00), Technologie minérale (271.00), Technologie du bâtiment et des travaux publics (221.00)
R14		Techniques cartographiques et géodésiques (230.00)
R15		Techniques policières (310.01), Techniques correctionnelles (310.02)
R16		Techniques administratives (410.00), Administration et coopération (413.01)
R17		Informatique (420.00)
R18		Techniques d'éducation spécialisée (351.00), Assistance sociale (388.00), Techniques de la documentation (393.00)
R19		Secrétariat (412.01)
R20		Arts plastiques (511.01), Musique (551.01), Esthétique de présentation (570.02), Aménagement d'intérieurs (570.03), Graphisme (570.06)

Tableau 2: Caractéristiques personnelles (suite)

Identifieur	Question	Modalité
*	Cegep	Collège de Lévis
*		F.-X. Garneau
*		La Pocatière
*		Lévis-Lauzon
*		Limoilou
*		Séminaire de Québec
*		Séminaire St-Augustin
*		Séminaire St-Georges
*		Sainte-Foy
*		Thetford-Mines

2. LA MÉTHODE D'ANALYSE DES CORRESPONDANCES

2.1 Objectif de la méthode

L'étude des relations entre les diverses variables de classement au test peut se faire par l'examen de tableaux croisant ces variables deux à deux. Bien que le nombre de tableaux à examiner soit assez restreint dans ce cas (une dizaine), il demeure difficile d'établir une vision globale des interrelations qui existent entre ces variables de classement. C'est l'objectif de la méthode d'analyse des correspondances que de nous fournir une image synthèse du problème étudié.

Chaque individu est ici décrit par son classement à toutes les parties du test et a même poids. Quant aux modalités de classement, leur poids est pris en compte par un choix particulier du critère de calcul de la distance entre les individus (distance pondérée entre profils). Enfin la méthode permet de situer sur un graphique l'ensemble des modalités de classement à chaque partie du test, de même que les individus.

2.2 Résumé de la méthode d'analyse factorielle des correspondances

Au départ, on se donne un tableau de données croisant deux caractéristiques I et J. I représente le nombre de lignes du tableau (nombre de modalités de la caractéristique I) et J le nombre de colonnes (nombre de modalités de la caractéristique J). Le tableau contient à l'intersection de la ligne i et de la colonne j le nombre k_{ij} d'individus ayant choisi à la fois les modalités i et j. En divisant l'effectif k_{ij} par l'effectif total k de toutes les cases du tableau, on obtient la donnée $f_{ij} = k_{ij} / k$, c'est-à-dire la fréquence des individus ayant les modalités i et j. La somme de tous les éléments d'une ligne du tableau global divisée par le grand total nous fournit la donnée f_i; c'est la fréquence des individus qui ont la modalité i, indépendamment de toute modalité j de J. De la même façon, le décompte des éléments d'une colonne du tableau divisé par le grand total nous fournit la donnée f_j; c'est la fréquence des individus ayant la modalité j, indépendamment de l'ensemble I.

Considérons la ligne i du tableau des fréquences f_{ij}, c'est-à-dire $f_{iJ} = \{f_{ij} \mid j \in J\}$. Si on divise cette ligne par son total f_i, on obtient

$$f_J^i = \left\{ \frac{f_{ij}}{f_i} \mid j \in J \right\},$$

appelé " profil de l'élément i sur l'ensemble des modalités de J (profil de la modalité i) ". De façon analogue, en divisant une colonne du tableau par son total f_j, on obtient

$$f_I^j = \left\{ \frac{f_{ij}}{f_j} \mid i \in I \right\},$$

appelé " profil de l'élément j sur l'ensemble des modalités de I (profil de la modalité j) ".

S'il y a indépendance entre les deux caractéristiques, c'est-à dire si toutes les modalités de I ont un profil similaire $\{f_j \mid j \in J\}$ ou si toutes les modalités de J ont un profil similaire $\{f_i \mid i \in I\}$, alors le tableau f_{IJ} est obtenu par une simple multiplication

$$f_{ij} = f_i \times f_j \ , \quad i \in I \quad \text{et} \quad j \in J.$$

Dans le cas contraire, nous cherchons à identifier les principaux facteurs de dépendance de ces caractéristiques. Pour cela, nous effectuons une analyse en composantes principales (ACP) sur l'ensemble des modalités de J ainsi qu'une ACP sur l'ensemble des modalités de I. La première ACP nous fournit un système d'axes que nous pouvons interpréter en situant l'ensemble des modalités de J sur ceux-ci; à chaque axe est associée une quantité (appelée valeur propre) qui exprime la part de la dispersion totale expliquée par celui-ci. La seconde ACP effectuée sur l'ensemble des modalités de I conduit aux mêmes valeurs propres et aux mêmes axes d'inertie qui sont maintenant interprétés par rapport à l'ensemble des modalités de I.

Le premier axe d'inertie u_1 correspond à la meilleure description possible de la position des points (de I ou J). Ceci revient à exiger que la somme des carrés des écarts de chaque point à cet axe soit minimale. Chaque axe additionnel répond au même critère et exige en plus d'être perpendiculaire à tous les axes précédents.

On obtient ainsi un mode de représentation des données qui fait ressortir les caractéristiques particulières aux deux thèmes étudiés. Un exposé détaillé de la méthode d'analyse factorielle des correspondances est présenté dans Bertier et Bouroche [2, chap. 8].

2.3 Application de la méthode à l'analyse des variables de classement au test

Considérons un individu particulier " i ". Pour chaque variable

de classement au test, le résultat de i est codé de la façon suivante:

$$r_j(i) = \begin{cases} 1 & \text{si } i \text{ est classé dans la modalité } j \\ 0 & \text{sinon,} \end{cases}$$

j parcourant l'ensemble des modalités de classement de la variable.

Exemple

	Combinatoire					Proportionnalité					Pourcentages			Énigmes			
	P-O	CI	CII	FI	FII	P-O	CI	CII	FI	FII	P-O	C	F	CI	CII	FI	FII
i	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	0

La lecture des résultats de l'individu i nous indique qu'il est classé concret II en combinatoire, concret I en proportionnalité, concret en pourcentages et concret I aux énigmes.

On obtient ainsi un tableau de données (noté X) d'environ 6 000 lignes et 17 colonnes; il croise l'ensemble des individus avec l'ensemble des variables de classement au test. Cependant, ce tableau peut présenter des problèmes de traitement étant donné sa grande dimension. De plus, l'intérêt de représenter les quelques 6 000 individus sur un graphique est discutable. Le problème du traitement informatique est contourné en soumettant à l'analyse des correspondances le tableau $B = {}^tX X$ * de dimensions 17 × 17. Selon Bastin et al. [1 , p. 312] l'analyse du tableau B fournit des facteurs proportionnels à ceux du tableau X. Il est donc équivalent du point de vue des associations entre les modalités de faire l'analyse sur B ou sur X.

En fait, le tableau B est un tableau symétrique qui consiste en la juxtaposition des seize tables de contingence que l'on obtient en croisant les variables de classement au test, deux à deux.

---

* ${}^tX$ désigne le tableau transposé.

Les caractéristiques personnelles, quant à elles, sont très variées. Ainsi, il nous a paru adéquat de les représenter en fonction des facteurs de l'analyse, ceux-ci étant déterminés uniquement au moyen des variables de classement.

3. RÉSULTATS OBTENUS

Au moyen de l'analyse des correspondances, nous cherchons à visualiser les liens qui peuvent exister entre les modalités de classement à chaque partie du test et entre les parties. Pour cela nous nous servons de graphiques dont les axes sont interprétés par rapport à l'ensemble des caractéristiques étudiées. La figure 1 nous fournit la représentation des modalités de chaque variable dans le plan des deux premiers axes issus de l'analyse factorielle. Les remarques qui suivent serviront à guider le lecteur dans l'interprétation des résultats:

a) La proximité de deux points du graphique indique que leurs profils se ressemblent; ces mêmes points peuvent cependant être plus distants lorsque représentés sur d'autres axes.

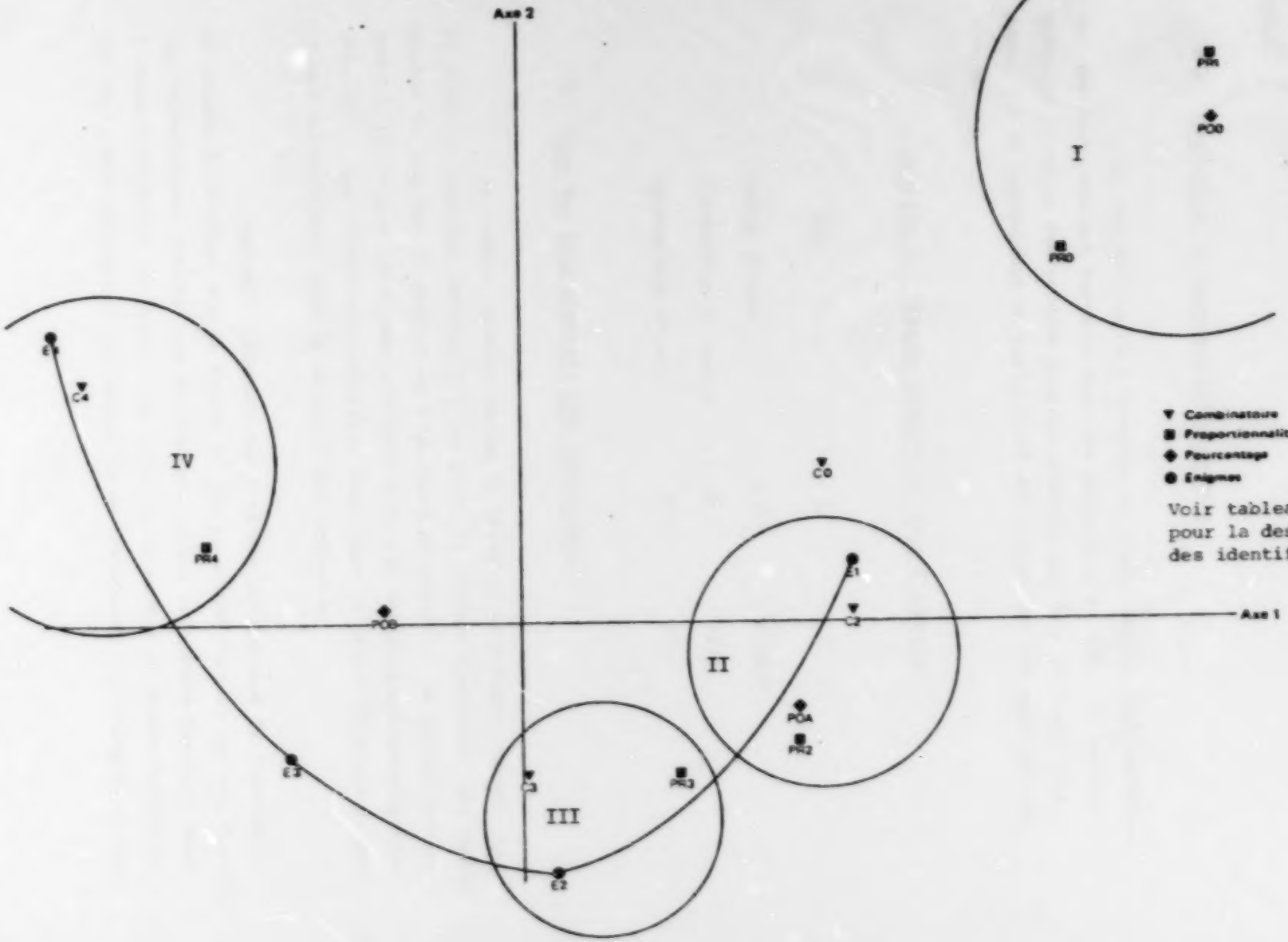
b) Des points éloignés ont des profils différents et feront éventuellement ressortir des oppositions particulières.

c) Des points proches du croisement des axes s'apparentent au profil moyen de l'ensemble.

d) Certains points utilisés dans une première analyse ont été écartés des analyses subséquentes (CO, PRO). Ceci a été fait afin de s'assurer que l'analyse n'avait pas été perturbée par ces éléments d'effectif faible. La méthode permet cependant de réintroduire ces points sur les cartes à une position déterminée à partir de leur profil et des facteurs obtenus. Ainsi, ces points ajoutent un caractère de vérification des résultats obtenus.

FIGURE 1

REPRESENTATION DES STADES PAR ANALYSE DES CORRESPONDANCES



3.1 Qualité de la représentation obtenue

La variabilité de l'ensemble de points analysé est expliquée par les deux premiers facteurs dans une proportion de 47%. Le tableau 3 présente la suite des valeurs propres associée aux deux premiers axes ainsi que le pourcentage de variabilité des points qui est expliqué par chacun.

Tableau 3: Valeurs propres et taux d'inertie

Axe	1	2
Valeur propre	0,277	0,0967
Pourcentage d'inertie	35,0	12,2
Pourcentage cumulé	35,0	47,2

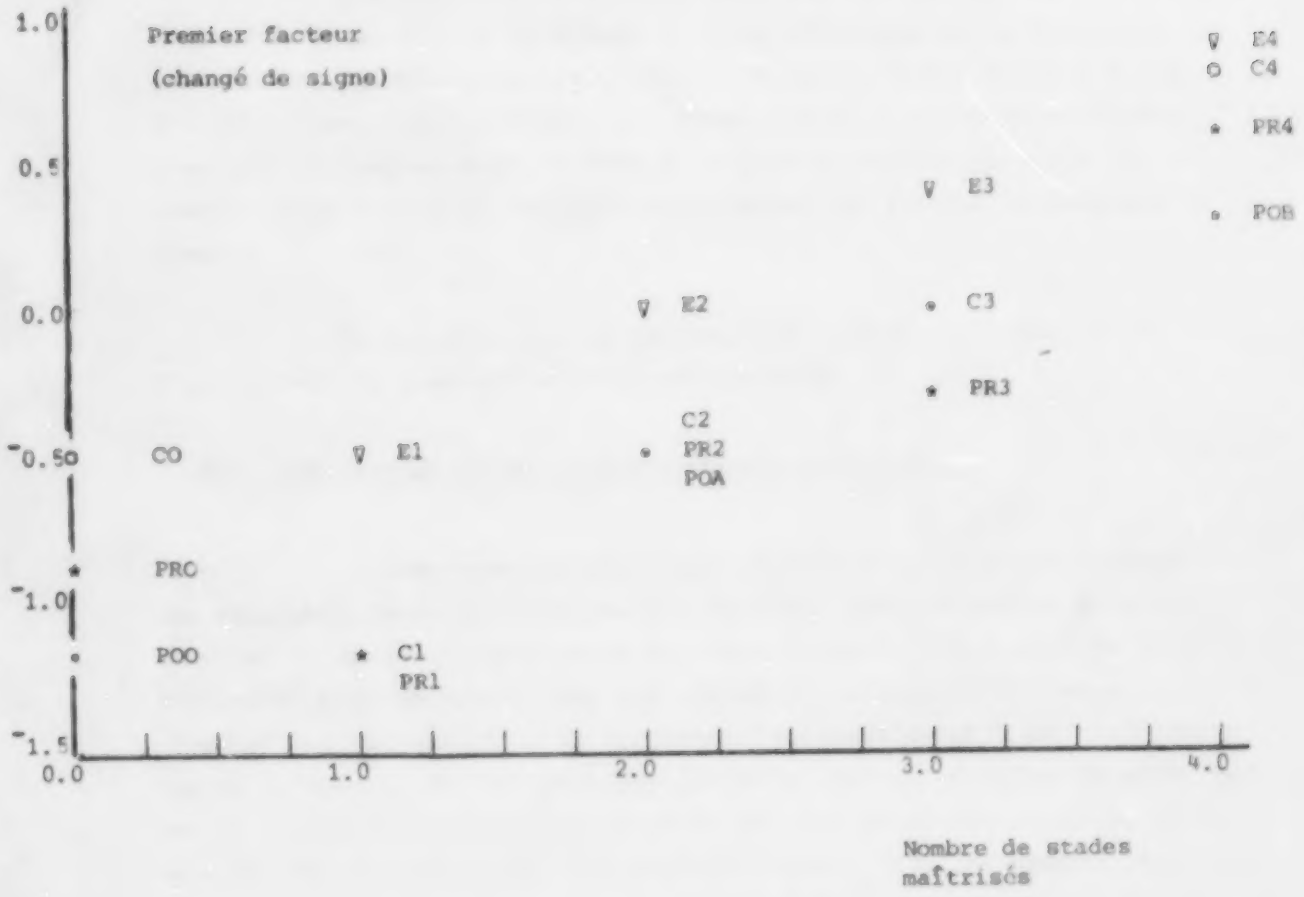
3.2 Plan des deux premiers axes factoriels

Le premier facteur oppose le stade le plus élevé (formel II) au stade le plus bas (concret I). De plus, il ordonne clairement les quatre stades du plus bas (à droite) au plus élevé (à gauche). Le second facteur quant à lui sépare les stades extrêmes (FII, CI) des stades intermédiaires (FI, CII). Les stades intermédiaires étant dans la région centrale du graphique caractérisent donc la majorité des individus.

On remarque également par l'allure parabolique de l'ensemble de points (voir tracé sur la figure 1) que le second facteur est une fonction approximativement quadratique du premier. Quant au premier facteur, nous l'avons reporté sur la figure 2 en fonction du nombre de stades maîtrisés que devraient représenter les modalités de classement (par exemple, l'indi-

Figure 2

Valeur des variables de classement sur le premier facteur vs le nombre de stades maîtrisés qu'elles représentent



Voir au tableau 1 pour la description des identifieurs.

vidu classé " E2 " devrait maîtriser les deux premiers stades des énigmes). On constate que pour chaque partie du test, le premier facteur est une fonction approximativement linéaire du nombre de stades maîtrisés.

Ces remarques sont faites pour préciser que la forme que prend l'ensemble de points sur la figure est caractéristique de la hiérarchie des modes de classement utilisés. En effet, selon Bastin et al. [1, pp. 383-385], lorsqu'une hiérarchie est présente dans un ensemble de données, l'analyse des correspondances fournit un premier facteur qui range les variables et un second qui est approximativement une fonction quadratique du premier.

On trouvera dans le rapport de M. Lagacé ([3 , chap. 8]), l'interprétation complète de cette configuration.

3.3 Représentation des caractéristiques personnelles

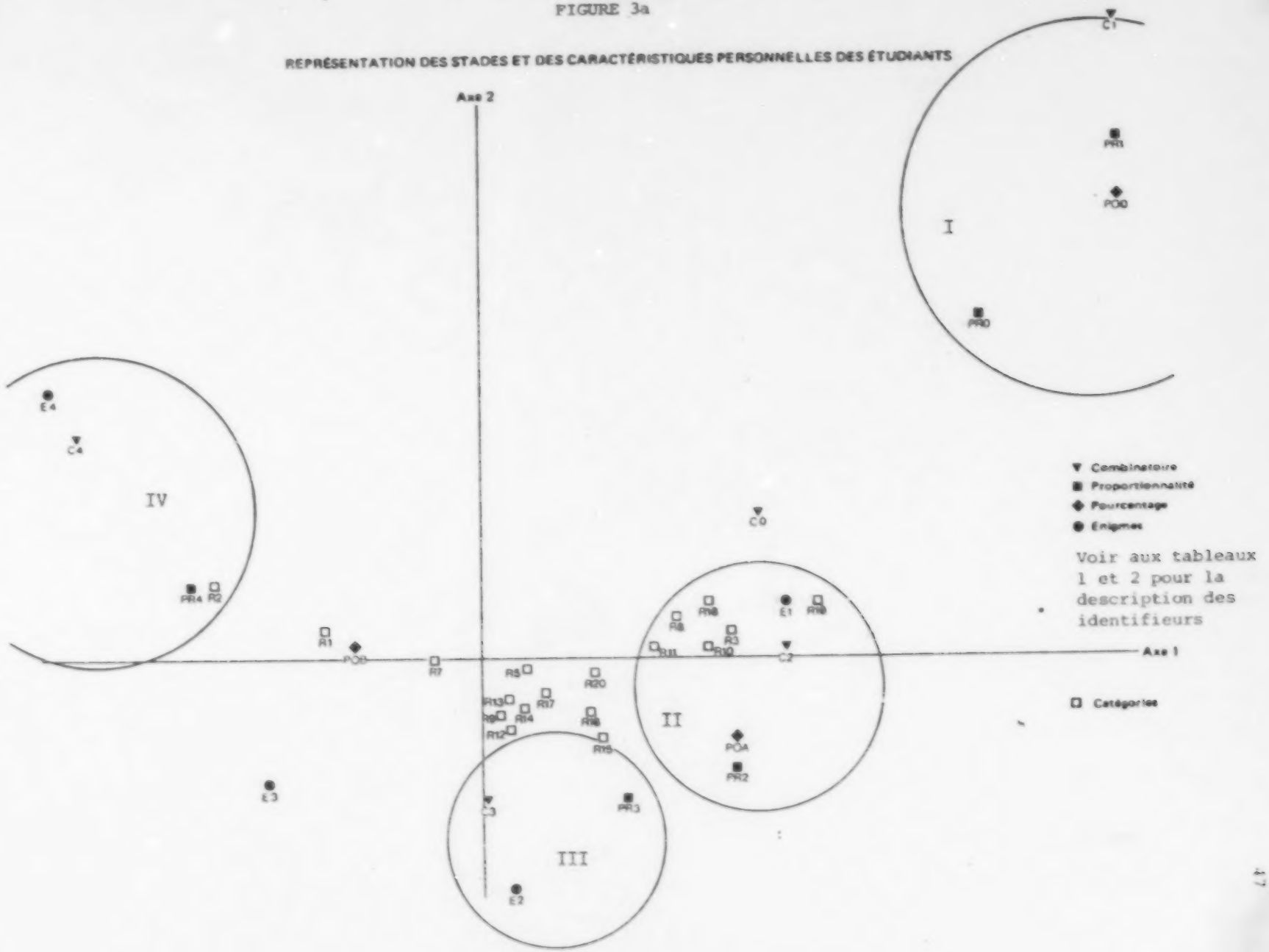
La configuration des stades peut être éclairée par l'ensemble des caractéristiques personnelles qui décrivent les individus. Nous relevons ici seulement quelques faits saillants laissant à M. Lagacé le soin d'élaborer plus longuement dans son rapport ([3 , chap. 8]). Ainsi, on remarque* que les catégories de programme " Sciences pures " (R2) et " Sciences de la santé " (R1) se détachent nettement dans la direction du stade formel II. Quant aux catégories R19, R18, R3, R10, R8 et R11, elles se retrouvent du côté des stades les plus faibles (concret II). On remarque également que tous les cegeps (*) sauf un se retrouvent autour du croisement des axes. Il ne semble également pas y avoir de différence importante entre les sexes (GAR, FIL) et les secteurs (GEN, PRF) étant donné leur position près du centre du graphique.

---

* Voir figures 3a à 3d.

FIGURE 3a

REPRÉSENTATION DES STADES ET DES CARACTÉRISTIQUES PERSONNELLES DES ÉTUDIANTS



Voir aux tableaux 1 et 2 pour la description des identifieurs

FIGURE 3b

REPRÉSENTATION DES STADES ET DES CARACTÉRISTIQUES PERSONNELLES DES ÉTUDIANTS

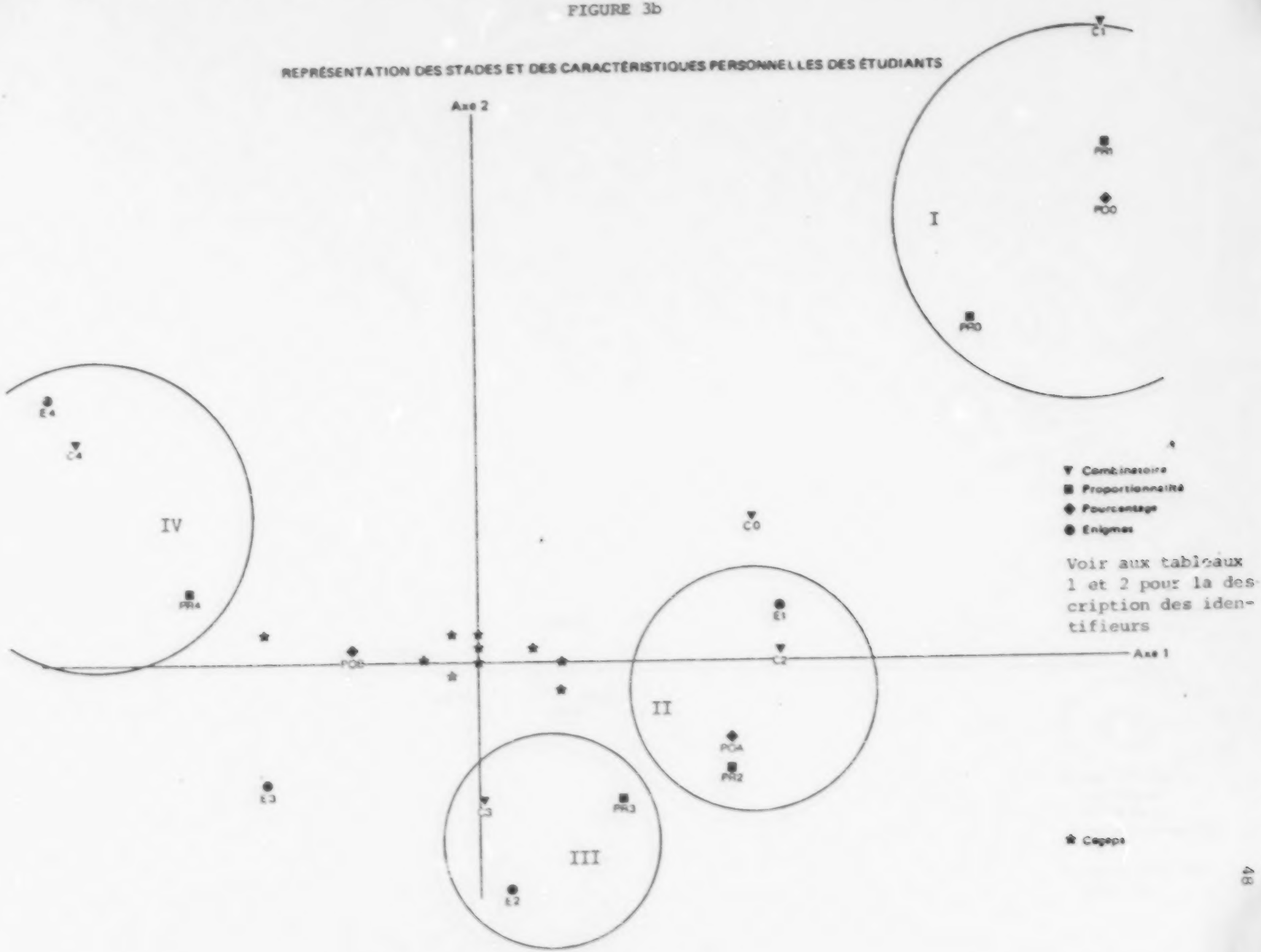


FIGURE 3c

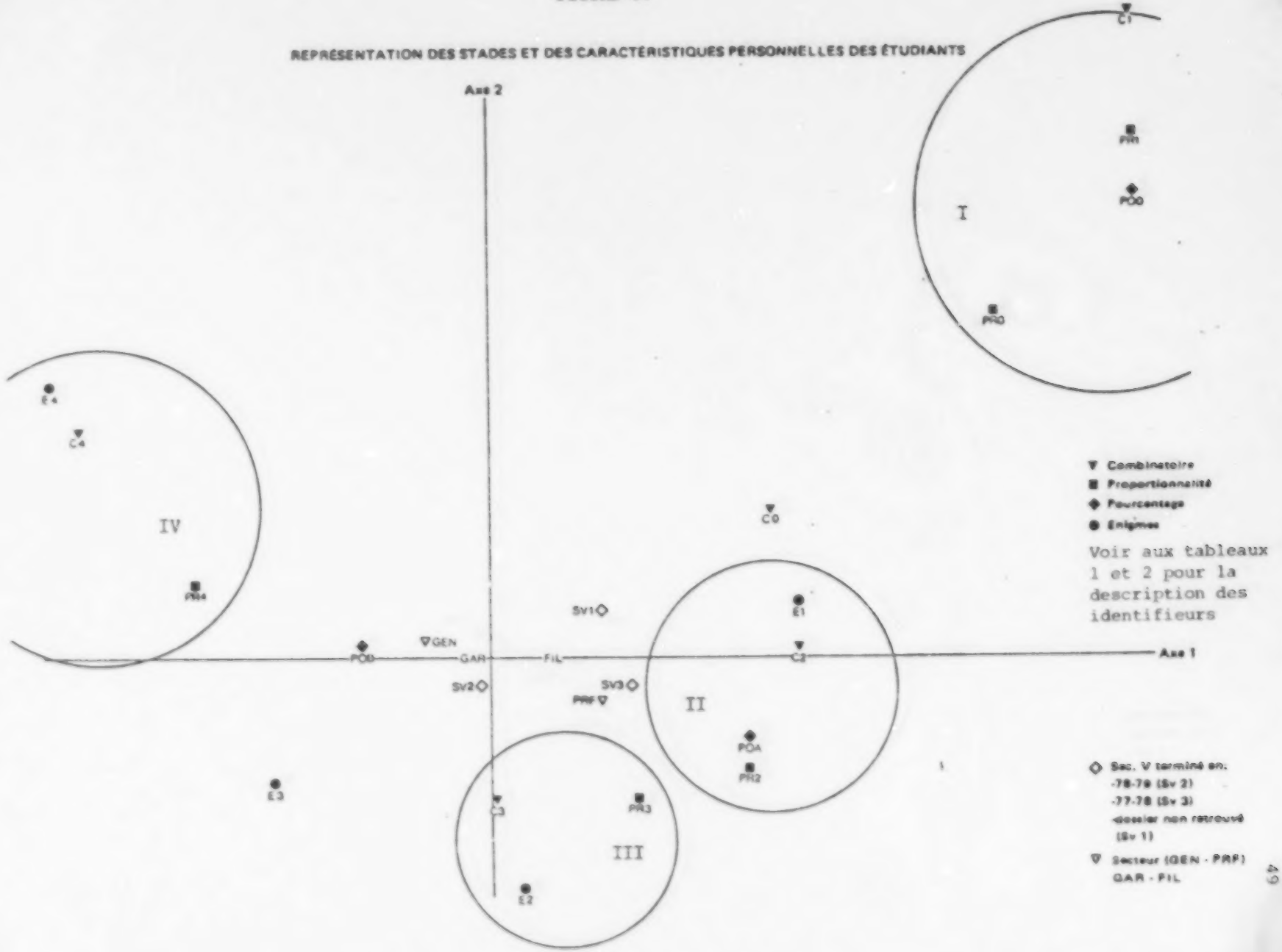
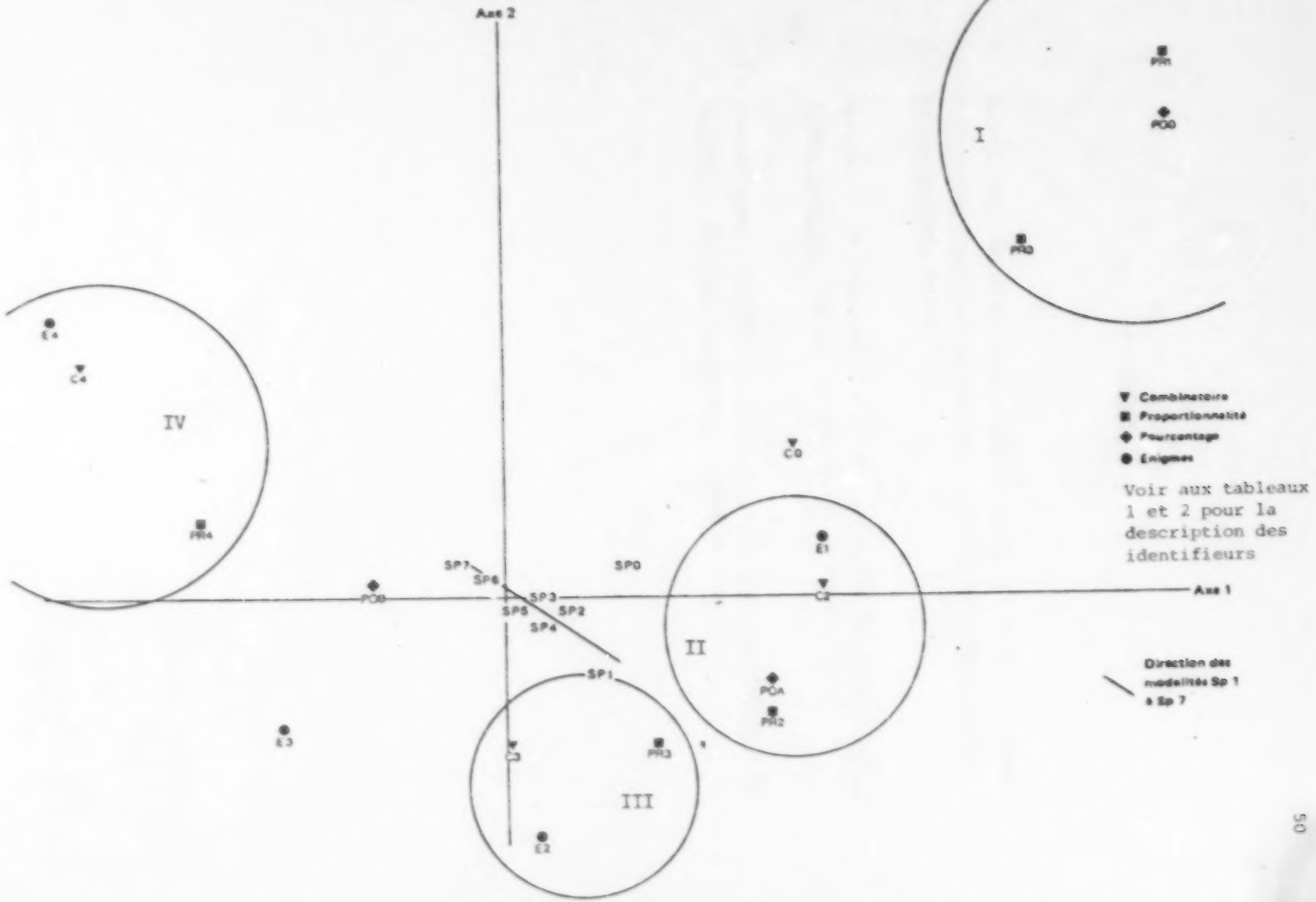

REPRÉSENTATION DES STADES ET DES CARACTÉRISTIQUES PERSONNELLES DES ÉTUDIANTS
Axe 2
Axe 1
C1
PIm
POD
I
PRD
E4
C4
IV
PR4
C0
E1
SV1
GEN
POD
GAR
FIL
C2
SV2
SV3
PRF
II
POA
PR2
E3
C3
PR3
III
E2
▼ Combinatoire
■ Proportionnalité
◆ Pourcentage
● Énigmes
Voir aux tableaux 1 et 2 pour la description des identifieurs
◇ Sec. V terminé en:
-78-79 (Sv 2)
-77-78 (Sv 3)
dossier non retrouvé (Sv 1)
▽ Secteur (GEN - PRF)
GAR - FIL

FIGURE 3d

REPRESENTATION DES STADES ET DES CARACTERISTIQUES PERSONNELLES DES ETUDIANTS



RÉFÉRENCES

[1] Bastin, Ch., Benzécri, J.-P., Bourgarit, Ch. et Cazes, P., (1980). Pratique de l'Analyse des données, 2: Abrégé théorique, Etudes de cas modèles, Dunod, Paris.

[2] Bertier, P. et Bouroche, J.M., (1977). Analyse des données multidimensionnelles, Presses universitaires de France, Paris.

[3] Torkia-Lagacé, Mirette, La pensée formelle chez les étudiants de collège I, objectif ou réalité ? Cegep de Limoilou.

V - ANALYSE DES CARACTÉRISTIQUES PERSONNELLES PAR UN MODÈLE LOG-LINÉAIRE

Dans le cadre du projet " La pensée formelle chez les étudiants de collège I ", nous voulons étudier l'influence des caractéristiques personnelles des étudiants sur les résultats au test.

L'étude simultanée de plusieurs variables qualitatives peut se faire en formant une table de contingence multidimensionnelle, chaque variable correspondant à une dimension de la table. L'analyse d'un tel ensemble de données est possible grâce aux modèles log-linéaires.

Dans un premier temps, pour mieux comprendre la démarche suivie dans ce type d'analyse, nous étudierons un modèle dans le cas d'une table à deux dimensions, puis le cas plus spécifique des tables à cinq dimensions sera étudié. Finalement, l'analyse des résultats et la recherche d'un modèle seront présentées dans la troisième partie de ce texte.

1. MODÈLE LOG-LINÉAIRE POUR UNE TABLE À DEUX DIMENSIONS

L'analyse des modèles log-linéaires a pour but d'obtenir une description des relations entre les variables de la table en vérifiant et ordonnant l'importance des interactions entre ces variables. L'analyse est basée sur l'ajustement d'un modèle hiérarchique[1] aux fréquences observées; c'est le logarithme des fréquences théoriques écrit comme une fonction additive d'effets principaux et d'interactions un peu de la même façon qu'un modèle d'analyse de variance.

Considérons une table I * J où l'on est en présence de deux variables A(j) et B(i) dont les décomptes x_{ij} sont illustrés dans le tableau 1.

---

1 Un modèle est hiérarchique s'il contient tous les effets d'ordre inférieur reliés aux effets d'ordre supérieur présents dans le modèle.

Tableau 1: Table de contingence à deux dimensions

B(i) \ A(j)							
	x_{11}	x_{12}		x_{1j}		x_{IJ}	$x_{1.}$
	x_{i1}			x_{ij}			$x_{i.}$
	x_{I1}					x_{IJ}	
	$x_{.1}$	$x_{.2}$		$x_{.j}$			N

On peut définir

$$x_{i.} = \sum_{j=1}^{J} x_{ij} \; ; \quad x_{.j} = \sum_{i=1}^{I} x_{ij} \; ; \quad N = \sum_{i=1}^{I} \sum_{j=1}^{J} x_{ij}.$$

Dans une table à deux dimensions, les estimés des fréquences théoriques de chaque cellule de la table sont de la forme suivante

$$m_{ij} = \frac{x_{i.}\, x_{.j}}{N} \tag{1}$$

sous l'hypothèse d'indépendance des rangées et des colonnes [1,3]. Prenant le logarithme naturel de l'expression (1), on obtient

$$\log m_{ij} = \log x_{i.} + \log x_{.j} - \log N. \tag{1'}$$

La forme additive du modèle (1') nous suggère d'exprimer le paramètre m_{ij} de la façon suivante

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} \tag{2}$$

où μ est la moyenne de tous les logarithmes des fréquences théoriques

$$\mu = \frac{1}{I\,J} \sum_{i=1}^{I} \sum_{j=1}^{J} \log m_{ij}.$$

L'expression $\mu + \mu_{1(i)}$ est l'effet principal dû à la variable A; c'est la moyenne des logarithmes des fréquences théoriques dans les j cellules à un niveau i de la première variable

$$\mu + \mu_{1(i)} = \frac{1}{J} \sum_{j=1}^{J} \log m_{ij}.$$

De même l'effet principal de la deuxième variable est

$$\mu + \mu_{2(j)} = \frac{1}{I} \sum_{i=1}^{I} \log m_{ij}.$$

Les termes $\mu_{1(i)}$ et $\mu_{2(j)}$ représentent les déviations de la moyenne générale μ et on pose

$$\sum_{i} \mu_{1(i)} = \sum_{j} \mu_{2(j)} = 0.$$

Au modèle exprimé en (2), on peut ajouter un terme d'interaction lorsqu'on veut considérer une dépendance entre les variables. On obtient alors le modèle

$$\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{1\,2(i,j)} \tag{3}$$

et l'on pose comme contraintes supplémentaires

$$\sum_{i=1}^{I} \mu_{1\ 2(i,j)} = \sum_{j=1}^{J} \mu_{1\ 2(i,j)} = 0.$$

On obtient des estimations du maximum de vraisemblance pour chacune de nos cellules sous un modèle hiérarchique par une méthode itérative décrite par Bishop, Fienberg et Holland [1].

La qualité de l'ajustement de notre modèle aux fréquences observées peut être vérifiée par la statistique du X^2 de Pearson

$$X^2 = \sum_{i,j} (x_{ij} - m_{ij})^2 / m_{ij}$$

ou par la statistique du rapport de vraisemblance

$$G^2 = 2 \sum_{i,j} x_{ij} \log (x_{ij} / m_{ij}).$$

Ces deux statistiques suivent une distribution du X^2 avec n - p degrés de liberté où n est le nombre de cellules et p le nombre de paramètres indépendants à estimer. La statistique du rapport de vraisemblance étant additive on pourra à partir de cette statistique vérifier un effet additionnel à un modèle. Ainsi pour vérifier l'effet de l'interaction $\mu_{1\ 2}$ dans le modèle exprimé en (3), on calculera pour chacun des deux modèles (2) et (3) la statistique G^2 et la différence

$$G^2_{(3)} - G^2_{(2)}$$

est une statistique qui permet de définir un test sur l'effet additionnel qui est l'interaction dans le modèle (3) conditionnellement aux effets dans

le modèle (2). Cette différence suit aussi une distribution du χ^2 avec comme degré de liberté la différence des degrés pour les statistiques de chacun des deux modèles. Cette propriété d'additivité de la statistique du rapport de vraisemblance ne tient pas pour la statistique du χ^2 de Pearson. C'est cette propriété qui fait que l'on utilisera principalement la statistique G^2 dans la recherche et l'analyse du modèle qui représentera le mieux nos observations.

2. LE MODÈLE LOG-LINÉAIRE POUR UNE TABLE À CINQ DIMENSIONS

Les modèles log-linéaires dans le cas des tables à plus de deux dimensions se construisent de la même façon et selon les mêmes principes que le cas à deux dimensions. Dans le cas d'un problème où l'on est en présence de cinq variables, le modèle saturé, c'est-à-dire le modèle où tous les effets possibles sont présents, est le suivant

$$\begin{aligned}\log m_{ijk\ell m} = {} & \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{4(\ell)} + \mu_{5(m)} \\ & + \mu_{12(i,j)} + \mu_{13(i,k)} + \mu_{14(i,\ell)} + \mu_{15(i,m)} \\ & + \mu_{23(j,k)} + \mu_{24(j,\ell)} + \mu_{25(j,m)} + \mu_{34(k,\ell)} \\ & + \mu_{35(k,m)} + \mu_{45(\ell,m)} + \mu_{123(i,j,k)} + \mu_{124(i,j,\ell)} \\ & + \mu_{125(i,j,m)} + \mu_{134(i,k,\ell)} + \mu_{135(i,k,m)} \\ & + \mu_{145(i,\ell,m)} + \mu_{234(j,k,\ell)} + \mu_{235(j,k,m)}\end{aligned} \tag{4}$$

$$+ \mu_{2\,4\,5\,(j,\ell,m)} + \mu_{3\,4\,5(k,\ell,m)} + \mu_{1\,2\,3\,4(i,j,k,\ell)}$$

$$+ \mu_{1\,2\,3\,5(i,j,k,m)} + \mu_{1\,2\,4\,5(i,j,\ell,m,)} + \mu_{1\,3\,4\,5(i,k,\ell,m)}$$

$$+ \mu_{2\,3\,4\,5\,(j,k,\ell,m)} + \mu_{1\,2\,3\,4\,5(i,j,k,\ell,m)}$$

avec le contraintes habituelles

$$\sum_i \mu_{1(i)} = \sum_j \mu_{2(j)} = \sum_k \mu_{3(k)} = \sum_\ell \mu_{4(\ell)} = \sum_m \mu_{5(m)} = 0$$

$$\sum_i \mu_{1\,2(i,j)} = \sum_j \mu_{1\,2(i,j)} = 0 \quad \text{et de même pour tous}$$

les autres termes d'ordre 2

$$\sum_i \mu_{1\,2\,3(i,j,k)} = \sum_j \mu_{1\,2\,3(i,j,k)} = \sum_k \mu_{1\,2\,3(i,j,k)} = 0$$

de même pour tous les autres termes d'ordre 3

$$\sum_i \mu_{1\,2\,3\,4(i,j,k,\ell)} = \sum_j \mu_{1\,2\,3\,4(i,j,k,\ell)} = \sum_k \mu_{1\,2\,3\,4(i,j,k,\ell)}$$

$$= \sum_\ell \mu_{1\,2\,3\,4(i,j,k,\ell)} = 0 \quad \text{ainsi que pour les autres}$$

termes d'ordre 4

$$\sum_i \mu_{1\,2\,3\,4\,5(i,j,k,\ell,m)} = \ldots = \sum_m \mu_{1\,2\,3\,4\,5(i,j,k,\ell,m)} = 0.$$

Nous ne considèrerons que les modèles hiérarchiques, c'est-à-dire les modèles où un effet d'ordre supérieur ne peut être présent à moins que tous les autres effets d'ordre inférieur reliés à cet effet supérieur ne

soient présents. On limite ainsi le nombre de modèles possibles à étudier mais nous sommes quand même devant une importante quantité de modèles.

Pour sélectionner le modèle qui s'ajuste le mieux à nos fréquences observées, nous utilisons la propriété d'additivité de G^2 la statistique du rapport de vraisemblance en comparant le modèle contenant tous les effets d'un certain ordre avec le modèle contenant un effet de moins.

Pour fixer les idées, si on enlève l'effet principal de la variable 1, le modèle devient alors[1]

$$\log m_{i,j,k,\ell,m} = \mu + \mu_2 + \mu_3 + \mu_4 + \mu_5 + \mu_{23} + \mu_{24} + \mu_{25} + \mu_{34} + \mu_{35} + \mu_{45} + \mu_{234} + \mu_{235} + \mu_{245} + \mu_{345} + \mu_{2345}. \quad (5)$$

La statistique $G^2_{(4)} - G^2_{(5)}$ nous permet de construire un test qui vérifiera si l'effet principal de la variable 1 est non nul. $G^2_{(4)}$ est la statistique du modèle représentée par l'expression (4) et $G^2_{(5)}$ est la statistique représentée par l'expression (5).

Le calcul de ces statistiques pour vérifier toutes les interactions de tout ordre nous permettra de connaître celles qui seront importantes dans le modèle. On pourra ainsi construire le modèle et qui nous informera sur les différents types de relations que l'on peut déceler entre les variables.

---

1 Afin d'alléger la notation, on omettra les parenthèses indiquant les niveaux i,j,k,ℓ et m des effets considérés. Ainsi, on écrira μ_2 plutôt que $\mu_2(j)$.

3. PRÉSENTATION DES RÉSULTATS

3.1 La table des observations

Les variables considérées sont le résultat à la partie énigme du test (B) [4], la catégorie d'étude (C), l'année de passation du D.E.S. (A), l'organisme de provenance (O) et finalement le sexe (S). La catégorie d'étude a été subdivisée en cinq: les étudiants de sciences santé, de sciences pures, le regroupement des autres catégories faisant partie du secteur général, le couple secrétariat et diététique et finalement le regroupement des autres catégories faisant partie du secteur professionnel. La variable organisme de provenance a été subdivisée en deux classes; le secteur privé et le secteur public. Le résultat à la partie énigme se subdivise en quatre modalités, soit ceux qui ont été classés concret I, concret II, formel I, formel II. Nous sommes donc en présence d'un ensemble de 5 327 individus que l'on peut représenter dans une table de contingence à cinq dimensions.

L'analyse du tableau a été produite à l'aide du programme BMDP3F [2]. Le tableau 2 présente les différents tests sur les effets principaux et sur les interactions de tout ordre. Ainsi à la première ligne du tableau, on présente les résultats du test qui portent sur la comparaison entre les deux statistiques

$$G^2 - G^{2'}$$

où G^2 est la statistique du modèle

$$\log m_{i,j,k,\ell,m} = \mu + \mu_B + \mu_S + \mu_O + \mu_A + \mu_C$$

et $G^{2'}$ est la statistique du modèle

Tableau 2: Test d'association partiel des facteurs[1]

Effet	D.L.	ΔG^2	Probabilité
B	3	1 043,71	0,0
S	1	0,48	0,4890
O	1	2 074,37	0,0
A	1	4 765,74	0,0
C	4	2 427,74	0,0
BS	3	5,90	0,1167
BO	3	2,00	0,5724
BA	3	1,75	0,6252
BC	12	616,25	0,0
SO	1	3,24	0,0719
SA	1	11,94	0,0005
SC	4	512,18	0,0
OA	1	1,68	0,1947
OC	4	90,21	0,0000
AC	4	97,07	0,0000
BSO	3	0,80	0,8498
BSA	3	1,41	0,7038
BSC	12	15,27	0,2270
BOA	3	1,11	0,7746
BOC	12	17,30	0,1385
BAC	12	21,59	0,0424
SOA	1	0,12	0,7243
SOC	4	7,23	0,1242
SAC	4	6,40	0,1711
OAC	4	1,93	0,7484
BSOA	3	3,31	0,3459
BSOC	12	4,81	0,9639
BSAC	12	5,77	0,9273
BOAC	12	8,72	0,9880
SOAC	4	0,94	0,9183
BSOAC	12	3,44	0,9916

1 Pour chaque effet d'ordre K, k=1, ..., 5, la statistique ΔG^2 est la différence entre l'ajustement du modèle hiérarchique contenant tous les effets d'ordre K et celui où l'on a enlevé l'effet étudié.

$$\log m_{i,j,k,\ell,m} = \mu + \mu_S + \mu_O + \mu_A + \mu_C.$$

On peut donc vérifier ici si l'inclusion de la variable résultat au test est significative dans un modèle qui ne comprendrait que les effets principaux des variables. Dans l'autre partie du tableau 2, on présente les résultats dans le cas où l'on considère un modèle avec tous les effets d'ordre 2. Ainsi, on peut vérifier que l'interaction résultat au test et le sexe n'a pas un apport significatif dans un modèle saturé d'ordre 2. L'observation du tableau 2 nous amène à conclure que le modèle qui ajustera le mieux notre table n'aura pas d'interaction d'ordre supérieur à 3.

On cherchera notre modèle en considérant l'interaction d'ordre 3, BAC et les interactions d'ordre 2, AC, OC, SC, SA, BC.

La recherche du modèle qui s'ajustera bien à nos observations nous amène maintenant à considérer les cinq modèles du tableau 3.

Tableau 3: Ajustement de quelques modèles hiérarchiques aux résultats

Modèle	Interactions du modèle	Degrés de liberté (d.ℓ.)	G^2	$P(\chi^2_{d.\ell} > G^2)$
1	BAC, BA, BC, AC, SA, OC, SC	109	91,66	0,885
2	BAC, BA, BC, AC, OC, SC	110	103,95	0,644
3	BAC, BA, BC, AC, SA, OC	113	630,78	0,0
4	BAC, BA, BC, AC, SA, SC	114	2 275,48	0,0
5	BA, BC, AC, SA, SC, OC	121	118,81	0,539

Tableau 4: Comparaison entre les principaux modèles hiérarchiques

Modèles comparées	ΔG^2	Degrés de liberté (d.ℓ.)	$P(\chi^2_{d.\ell} > \Delta G^2)$
5 - 1	27,15	12	0,007
4 - 1	2 183,82	5	0,0
3 - 1	539,12	4	0,0
2 - 1	12,29	1	0,0

Le tableau 4 nous montre différentes comparaisons qui nous permettront de connaître l'importance des interactions dans le modèle. Ainsi, la statistique $G^2_5 - G^2_1$ nous indique que l'interaction d'ordre 3, BAC est importante dans le modèle qui reproduira notre table de contingence. Les trois autres comparaisons nous indiquent que les interactions du second ordre, SA, OC, SC sont également nécessaires pour expliquer le phénomène et reproduire adéquatement notre table.

On conclut donc que le meilleur modèle pour représenter les relations entre les différentes variables est

BAC, SA, OC, SC

qui s'écrit également comme

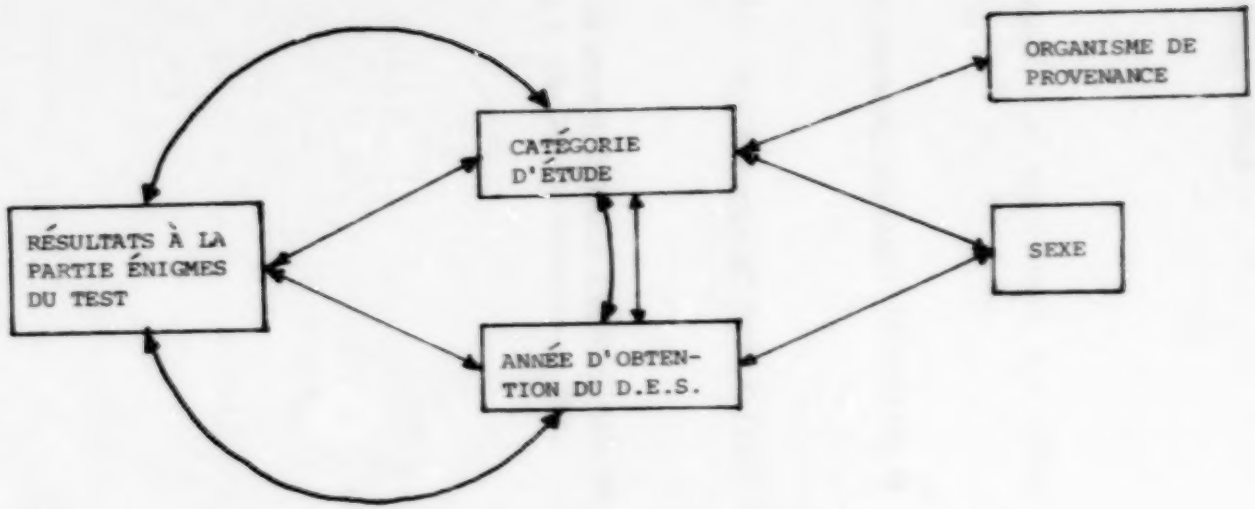
$$\log m_{i,j,k,\ell,m} = \mu + \mu_B + \mu_A + \mu_C + \mu_O + \mu_S + \mu_{BA} + \mu_{BC} + \mu_{AC} + \mu_{SA} + \mu_{SC} + \mu_{OC} + \mu_{BAC}.$$

Une illustration de ce modèle est donnée dans la figure 1. Les flèches entre les variables indiquent une interaction d'ordre 2. On peut remarquer immédiatement qu'il n'y a pas d'interaction entre l'organisme de provenance et les résultats au test, de même qu'entre le résultat au test et le sexe.

Le cercle sur la figure représente l'interaction d'ordre 3 entre les variables résultat au test, catégorie d'étude et année d'obtention du D.E.S. Cette interaction nous dit qu'il existe un lien significatif entre les trois variables, malgré le fait que la statistique de l'interaction entre le résultat au test et l'année d'obtention du D.E.S. ne soit pas significative dans le contexte du tableau.

REPRÉSENTATION DES INTERACTIONS DÉCELÉES PAR UN

MODÈLE LOG-LINÉAIRE



BIBLIOGRAPHIE

[1] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1976). Discrete Multivariate Analysis. 2nd Ed., MIT Press.

[2] BMDP-77. Biomedical Computer Programs. University at California Press.

[3] Fienberg, S.E. (1976). The Analysis of Cross-Classified Categorical Data. MIT Press.

[4] Torkia-Lagacé, M. La pensée formelle des étudiants de collège I; objectif ou réalité ? Collège de Limoilou. 1981.

VI - ANALYSE FACTORIELLE DES CORRESPONDANCES ENTRE LE CLASSEMENT AU TEST ET LES RÉSULTATS SCOLAIRES

A l'automne 1979, les étudiants de collégial I de la région 03 ont été soumis à un test constitué de quatre parties (combinatoire, proportions, pourcentages, énigmes). Chaque individu est caractérisé par son classement à chacune des parties du test; il s'agit d'un classement en quatre stades identifiés (concret I, concret II, formel I, formel II). Par la suite, on a adjoint au classement de chaque individu un ensemble de résultats scolaires obtenus dans douze cours du secondaire et six cours du collégial. Nous avons utilisé l'analyse factorielle des correspondances pour étudier la relation qu'il peut y avoir entre le classement au test et les résultats scolaires.

1. DESCRIPTION DÉTAILLÉE DES DONNÉES

Les deux ensembles de caractéristiques suivants sont considérés. Le premier est constitué par l'ensemble des modalités de classement à chaque partie du test et est présenté au tableau 1 avec un identifieur pour chaque modalité. Dans les trois premières parties du test, un individu ne maîtrisant aucun des quatre stades était classé " pré-opératoire " par défaut. Dans la quatrième partie, l'individu était classé " concret I " par défaut s'il ne maîtrisait aucun des trois stades suivants. Par ailleurs, dans la partie " pourcentages ", les individus étaient classés en fonction de deux stades seulement (concret, formel).

Tableau 1: Variables de classement au test

Identifieur	Partie	Stade
C0	Combinatoire	Pré-opératoire
C1	"	Concret I
C2	"	Concret II
C3	"	Formel I
C4	"	Formel II
PR0	Proportions	Pré-opératoire
PR1	"	Concret I
PR2	"	Concret II
PR3	"	Formel I
PR4	"	Formel II
P00	Pourcentages	Pré-opératoire
P0A	"	Concret
P0B	"	Formel
E1	Énigmes	Concret I
E2	"	Concret II
E3	"	Formel I
E4	"	Formel II

Le second ensemble concerne dix-huit cours du secondaire et du collégial I. Afin de pouvoir traiter ces renseignements avec ceux du premier ensemble (test), nous avons dû découper les résultats scolaires en quartiles. Le tableau 2 présente l'ensemble des variables de résultats scolaires avec les étiquettes qui les identifient par la suite.

Tableau 2: Résultats scolaires

Identifieur	Cours
1C	Chimie 442
2C	" 462
3C	" 562
1F	Français 512
2F	" 522
3F	" 532
1M	Mathématiques 522
2M	" 532
1P	Physique 422
2P	" 432
3P	" 522
4P	" 552
CB	Collégial, biologie
CC	" , chimie
CF	" , français
CH	" , philosophie
CM	" , mathématiques
CP	" , physique

A ces identifieurs s'ajoute un troisième caractère défini de la façon suivante:

0 : n'a pas suivi le cours
1 : étudiant classé dans le 1er quartile
2 : étudiant classé dans le 2^{e} quartile
3 : étudiant classé dans le 3^{e} quartile
4 : étudiant classé dans le 4^{e} quartile

2. LA MÉTHODE D'ANALYSE DES CORRESPONDANCES

2.1 Objectif de la méthode

L'étude des relations entre les résultats au test et les résultats scolaires s'appuie sur un ensemble de données assez vaste. En effet, il y a d'une part les quatre parties du test à considérer, d'autre part dix-huit cours aux niveaux secondaire et collégial. Une première approche qui concerne l'examen des relations entre ces deux groupes de caractéristiques, nous amène à scruter soixante-douze tables de contingence. L'information ainsi obtenue demeure difficile à synthétiser et est d'une consultation aride. Il apparaît donc intéressant d'utiliser une méthode d'analyse qui fournisse dans un premier temps une vision globale du problème étudié. Cette méthode tiendra compte d'un certain nombre d'interrelations qui existent entre les caractéristiques.

Au moyen de l'analyse factorielle des correspondances, il est possible d'étudier la relation entre les variables de classement au test (tableau 1) et les résultats scolaires (tableau 2). De plus, on a l'avantage de pouvoir situer sur un même graphique les deux ensembles de caractéristiques. La méthode tient également compte du poids des observations étant donné que l'analyse est effectuée sur des profils; on peut ainsi comparer des observations de poids différents. Enfin, par le choix particulier du critère de calcul de la distance entre les observations (distance pondérée entre profils), les résultats obtenus dépendent peu de la nomenclature choisie pour la construction des classes; en effet si deux colonnes du tableau des données sont proportionnelles, on pourra les réunir en une seule classe correspondant aux effectifs des deux classes réunies.

2.2 Résumé de la méthode d'analyse factorielle des correspondances

Au départ, on se donne un tableau de données croisant deux

caractéristiques I et J. I représente le nombre de lignes du tableau (nombre de modalités de la caractéristique I) et J le nombre de colonnes (nombre de modalités de la caractéristique J). Le tableau contient à l'intersection de la ligne i et de la colonne j le nombre k_{ij} d'individus ayant choisi à la fois les modalités i et j. En divisant l'effectif k_{ij} par l'effectif total k de toutes les cases du tableau, on obtient la donnée $f_{ij} = k_{ij} / k$, c'est-à-dire la fréquence des individus ayant les modalités i et j. La somme de tous les éléments d'une ligne du tableau global divisée par le grand total nous fournit la donnée f_i; c'est la fréquence des individus qui ont la modalité i, indépendamment de toute modalité j de J. De la même façon, le décompte des éléments d'une colonne du tableau divisé par le grand total nous fournit la donnée f_j; c'est la fréquence des individus ayant la modalité j, indépendamment de l'ensemble I.

Considérons la ligne i du tableau des fréquences f_{ij}, c'est-à-dire $f_{iJ} = (f_{ij} \mid j \in J)$. Si on divise cette ligne par son total f_i, on obtient

$$f_J^i = \left\{ \frac{f_{ij}}{f_i} \mid j \in J \right\},$$

appelé " profil de l'élément i sur l'ensemble des modalités de J (profil de la modalité i) ". De façon analogue, en divisant une colonne du tableau par son total f_j, on obtient

$$f_I^j = \left\{ \frac{f_{ij}}{f_j} \mid i \in I \right\},$$

appelé " profil de l'élément j sur l'ensemble des modalités de I (profil de la modalité j) ".

S'il y a indépendance entre les deux caractéristiques, c'est-à-dire si toutes les modalités de I ont un profil similaire $\{f_j \mid j \in J\}$ ou si toutes les modalités de J ont un profil similaire $\{f_i \mid i \in I\}$, alors le tableau f_{IJ} est obtenu par une simple multiplication

$$f_{ij} = f_i \times f_j , \quad i \in I, \; j \in J.$$

Dans le cas contraire, nous cherchons à identifier les principaux facteurs de dépendance de ces caractéristiques. Pour cela, nous effectuons une analyse en composantes principales (ACP) sur l'ensemble des modalités de J ainsi qu'une ACP sur l'ensemble des modalités de I. La première ACP nous fournit un système d'axes que nous pouvons interpréter en situant l'ensemble des modalités de J sur ceux-ci; à chaque axe est associée une quantité (appelée valeur propre) qui exprime la part de la dispersion totale expliquée par celui-ci. La seconde ACP effectuée sur l'ensemble des modalités de I conduit aux mêmes valeurs propres et aux mêmes axes d'inertie qui sont maintenant interprétés par rapport à l'ensemble des modalités de I.

Le premier axe d'inertie u_1 correspond à la meilleure description possible de la position des points (de I ou J). Ceci revient à exiger que la somme des carrés des écarts de chaque point à cet axe soit minimale. Chaque axe additionnel répond au même critère et exige en plus d'être perpendiculaire à tous les axes précédents.

On obtient ainsi un mode de représentation des données qui fait ressortir les caractéristiques particulières aux deux thèmes étudiés. Un exposé détaillé de la méthode d'analyse factorielle des correspondances est présenté dans Bertier et Bouroche [1 , chap. 8].

2.3 Tableau des données analysées

Dans la présente étude, plusieurs caractéristiques sont prises

en considération. On considère d'une part les quatre parties du test et d'autre part les dix-huit variables de cours. Avec chaque partie et chaque variable de cours, nous construisons un tableau de données qui contient à la croisée de la ligne i (modalité de classement d'une partie) et de la colonne j (modalité d'une variable de cours) le nombre k_{ij} de personnes classées aux niveaux i de la partie et j du cours. Les quatre parties et les dix-huit cours interviennent dans la construction d'un grand tableau qui constitue une juxtaposition des soixante-douze tableaux de contingence réalisés sur le même échantillon. Les propriétés de l'analyse factorielle des correspondances sur de tels tableaux ont été étudiées par A. Leclerc [2].

3. RÉSULTATS OBTENUS

L'analyse des correspondances sert ici à mettre en relief les caractéristiques particulières aux sujets étudiés. Pour ce, nous nous servons de graphiques et nous prenons avantage du fait que nous pouvons situer sur un même plan à la fois l'ensemble des variables de classement au test et celui des cours. Les remarques suivantes s'appliquent quant à l'interprétation des résultats de l'un ou l'autre ensemble. Ainsi pour un ensemble de points donné (par exemple, les points cours):

a) La proximité de deux points du graphique indique que leurs profils se ressemblent; ces mêmes points peuvent cependant être plus distants lorsque représentés sur d'autres axes.

b) Des points éloignés ont des profils différents et feront éventuellement ressortir des oppositions particulières.

c) Des points proches du croisement des axes s'apparentent au profil moyen de l'ensemble.

De plus, on peut établir des relations entre les deux ensembles de données (points cours et variables de classement au test) de la façon

suivante:

d) Lorsque l'on évalue des proximités sur un ensemble (par exemple, les variables de classement), les points de l'autre ensemble (les cours) jouent le rôle de variables et on s'intéressera aux angles que ces points font avec les facteurs.

e) On peut étudier la position d'une modalité d'un ensemble par rapport à l'ensemble des modalités de l'autre. Cependant, on peut difficilement tirer des conclusions sur la proximité de deux points n'appartenant pas au même ensemble en ne tenant compte que de la position de ces deux points.

Enfin, certains points utilisés dans une première analyse peuvent être écartés par la suite. Ce procédé est employé lorsque les premiers résultats laissent soupçonner que les points en question ont un effet perturbateur. Plusieurs causes sont possibles (par exemples, effectif très faible ou très élevé, contribution très forte à la variance d'un facteur). La méthode utilisée nous permet cependant de réintroduire ces points sur les graphiques à une position déterminée à partir de leur profil et des facteurs obtenus. Dans le cas présent, les points CO et PRO, d'effectif très faible, n'ont pas participé à la construction des axes.

3.1 Qualité de la représentation obtenue

La variabilité des ensembles de points analysés est expliquée par les deux premiers facteurs dans une proportion de 96%; le tableau 3 présente la suite des valeurs propres associée aux deux premiers axes d'inertie ainsi que le pourcentage de dispersion des points qui est expliqué par chacun.

Tableau 3: Valeurs propres et taux d'inertie

Axe	1	2
Valeur propre	0,042368	0,002155
Pourcentage d'inertie	91,2	4,6
Pourcentage cumulé	91,2	95,8

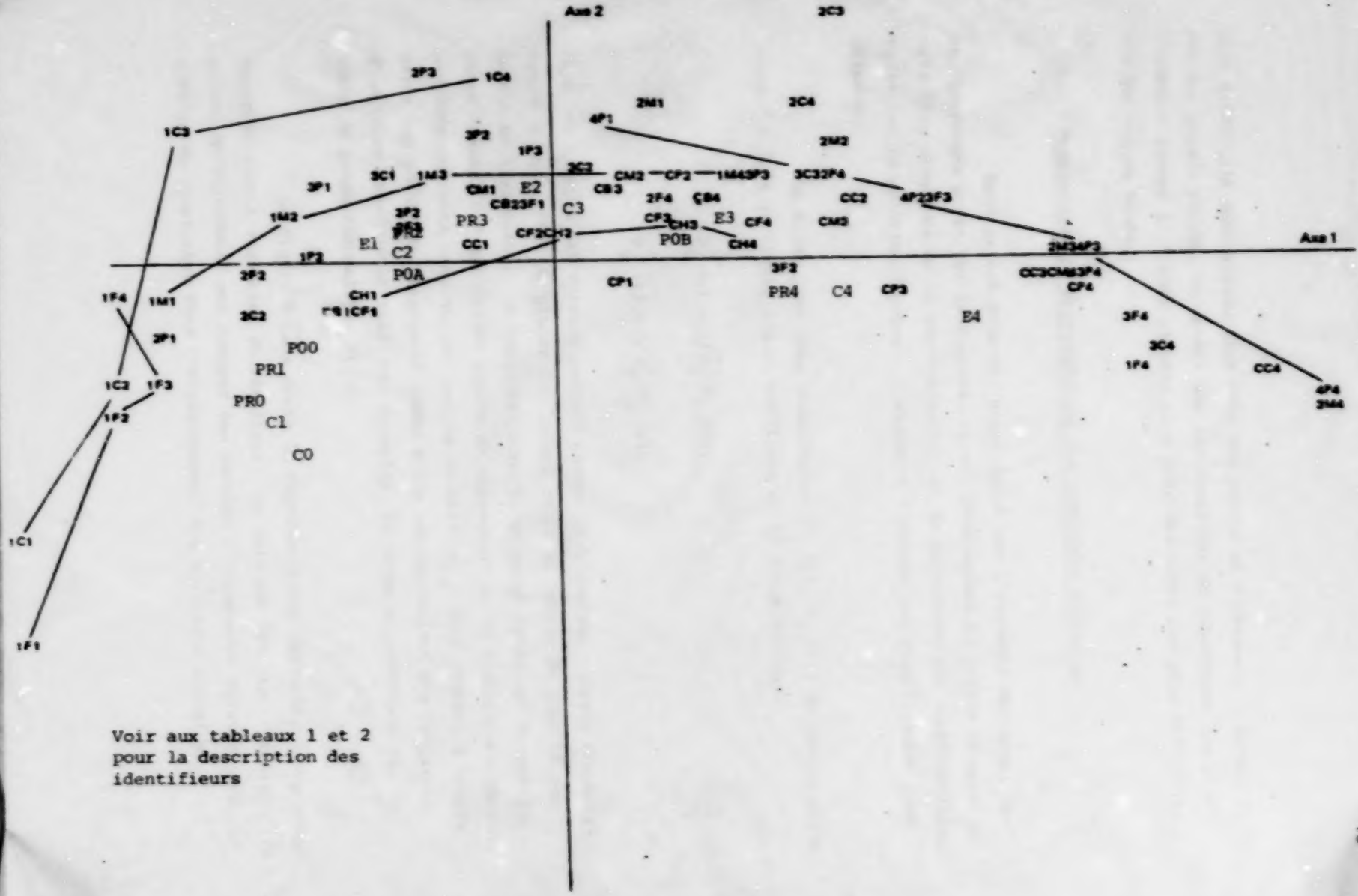
3.2 Plan des deux premiers axes factoriels

La figure 1 donne la représentation des variables du test et des cours dans le plan des deux premiers facteurs issus de l'analyse. Les modalités de classement de chaque partie du test sont ordonnées de gauche à droite, du stade le plus bas au stade le plus élevé. Sur le second axe, les stades intermédiaires de chaque partie s'opposent aux stades extrêmes. Quant aux points cours, le premier facteur les ordonne selon les quartiles (1^{er} à gauche au 4^{e} à droite). De plus, les cours situés dans la partie gauche sont les cours considérés plus faciles tandis que ceux à droite sont considérés les plus difficiles. Par rapport à l'axe 2, on retrouve en haut les modalités les plus élevées (3^{e}, 4^{e} quartiles) des cours faciles et les modalités les plus basses (1^{er}, 2^{e} quartiles) des cours plus difficiles; vers le bas de l'axe 2, on retrouve les modalités les plus faibles des cours les plus faciles (1^{er}, 2^{e} quartiles dans la partie inférieure gauche) et les modalités les plus fortes des cours plus difficiles (3^{e}, 4^{e} quartiles). Il faut cependant mentionner une exception à ce comportement, le cours de français 512 (IF) dont toutes les modalités se retrouvent dans la partie inférieure gauche du graphique.

La représentation simultanée des points cours et des points de classement (figure 1) nous permet d'étudier chaque ensemble de points par rapport à l'autre. On constate par exemple que les cours considérés les

FIGURE 1

REPRESENTATION DES COURS PAR L'ANALYSE DES CORRESPONDANCES



Voir aux tableaux 1 et 2
pour la description des
identifieurs

plus difficiles sont situés plus près des points de classement " formel II " que des autres points; ou encore que les modalités de classement les plus élevées " formel II " sont situées plus près des cours les plus difficiles que des autres cours.

3.3 Représentation barycentrique des résultats scolaires

Dans le cas présent, étant donné que l'ensemble des modalités de classement n'est pas trop grand, il est intéressant de mettre de côté le caractère simultané de la représentation et de considérer une représentation telle que la relation " cours - classement " prenne une signification plus précise.

En effet, les deux coordonnées (F_1 (i), F_2 (i)) de chaque point cours " i " de la figure 1 sont modifiées de la façon suivante:

$$F_1'(i) = \sqrt{\lambda_1}\, F_1(i),$$

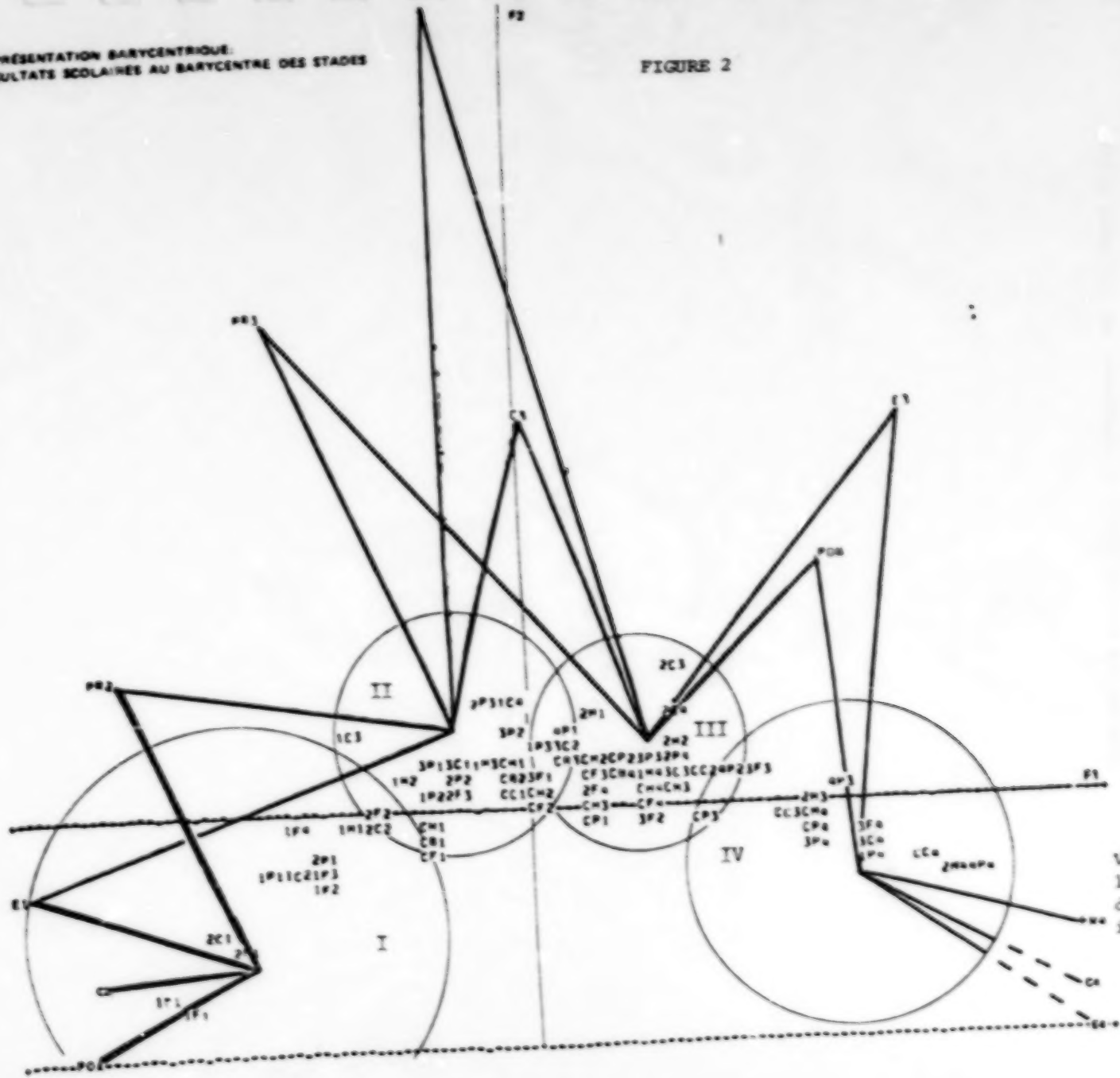
$$F_2'(i) = \sqrt{\lambda_2}\, F_2(i),$$

où λ_1 et λ_2 sont les valeurs propres issues de l'analyse. Cette transformation a pour effet de placer les points cours au centre de gravité des points de classement. On considère donc qu'un point cours est en quelque sorte " composé " des divers points de classement et on établit une moyenne pondérée de cette composition (centre de gravité). Cela revient à considérer les points de classement comme s'ils représentaient les résultats d'un cours très particulier (par exemples, le cours de combinatoire, le cours de proportionnalité, etc.).

La figure 2 nous donne la représentation discutée. Nous avons encerclé quatre ensembles de résultats. On voit par les traits reliant les points de classement aux centres des cercles l'importance relative des variables de classement dans l'établissement des résultats scolaires.

REPRÉSENTATION BARYCENTRIQUE
RÉSULTATS SCOLAIRES AU BARYCENTRE DES STADES

FIGURE 2



Voir aux tableaux 1 et 2 pour la description des identifieurs

Dans le rapport de M. Lagacé [3 , chap. 9], une interprétation complète des résultats de cette figure est donnée.

RÉFÉRENCES

[1] Bertier, P. et Bouroche, J.M., (1977), Analyse des données multidimensionnelles, Presses universitaires de France, Paris.

[2] Leclerc, A. (1975), l'Analyse des correspondances sur juxtaposition de tableaux de contingence, Revue de statistique appliquée, vol. 23, no 3, pp. 5-16.

[3] Torkia-Lagacé, M., La pensée formelle chez les étudiants de collège I, objectif ou réalité ? Cegep de Limoilou.

VII - ANALYSE DES RÉSULTATS SCOLAIRES SELON LE CLASSEMENT À LA PARTIE ÉNIGMES DU TEST PAR DES ANALYSES DE LA VARIANCE

Dans le cadre du projet " La pensée formelle chez les étudiants de collège I ", on veut comparer les résultats scolaires au secondaire et au collégial des étudiants selon leur classement à la partie énigmes du test. On traitera donc les résultats scolaires à l'aide de l'analyse de variance à classification simple. Cette analyse nous permettra de savoir s'il y a des différences significatives entre les résultats à un cours donné selon qu'ils se classent concret I, concret II, formel I ou formel II à la partie énigmes du test. Les résultats des cours au collégial sont les notes que les étudiants ont obtenues dans leur groupe respectif moins la moyenne de ce groupe. Nous avons effectué cette transformation dans le but d'éliminer les différences dans les résultats dues à des modes d'évaluation différents ainsi qu'à des barèmes de compilation différents. Pour les cours au secondaire, nous avons conservé la note brute puisqu'ils ont tous été soumis à un examen commun du ministère.

Dans un premier temps, nous allons exposer brièvement la méthode d'analyse de variance ainsi que les possibilités de comparaison entre les différents stades atteints. Finalement, nous regarderons les résultats obtenus pour les cours au secondaire et les cours au collégial.

1. L'ANALYSE DE VARIANCE

Le plan à classification simple est un plan d'expérience à un seul facteur (disons à K niveaux). Le contexte est le suivant: soit K populations indépendantes avec moyennes respectives $\mu_1, \mu_2, \ldots, \mu_K$ et variance commune σ^2. Les paramètres $\mu_1, \mu_2, \ldots, \mu_K$ et σ^2 sont inconnus. On prend des échantillons aléatoires indépendants de tailles respectives $n_1, n_2, \ldots, n_K$, c'est-à-dire $X_{11}, X_{12}, \ldots, X_{1n_1}$ de la première population et de même pour les K - 1 autres populations et on veut tester l'hypothèse de l'égalité des K moyennes μ_i.

Le modèle mathématique qui soutient cette analyse est de la forme

$$X_{ij} = \mu_i + \varepsilon_{ij} \qquad \begin{matrix} i = 1, 2, \ldots, K \\ j = 1, 2, \ldots, n_i \end{matrix}$$

avec

$$E(\varepsilon_{ij}) = 0$$

$$E(\varepsilon_{ij}\, \varepsilon_{i'j'}) = \sigma^2 \qquad i = i' \,,\ j = j'$$

$$= 0 \qquad \text{autrement.}$$

On estimera la variance de la population de deux façons et on les comparera pour vérifier notre hypothèse d'égalité des moyennes.

On calcule d'abord une variance pondérée pour les groupes

$$s_p^2 = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} X_{ij}^2 - \sum_{i=1}^{K} (X_{i.}^2 / n_i)}{\sum n_i - K}$$

et on calcule la variance de la moyenne directement

$$s_M^2 = \sum_{i=1}^{K} \frac{n_i \, (\bar{X}_{i.} - \bar{X}_{..})^2}{K - 1}$$

Le rapport de ces deux estimateurs de la variance nous permettra de construire le test qui vérifiera l'hypothèse suivant:

$$H_o : \mu_1 = \mu_2 = \ldots = \mu_K$$

versus $H_1 : \mu_i \neq \mu_j$ pour au moins une valeur de i et de j.

On rejettera cette hypothèse d'égalité des moyennes si le rapport S_M^2 / S_p^2 excède la valeur critique d'une loi de Fisher avec K - 1 et $\sum_{i=1}^{K} n_i$ - K degrés de liberté.

L'analyse de variance n'est que la première étape dans l'analyse des résultats. L'étape suivante est d'examiner les moyennes des groupes et de vérifier des hypothèses d'égalité entre les moyennes de certains groupes. On pourra effectuer ces comparaisons en considérant des contrastes qui sont des combinaisons linéaires des moyennes des groupes. Soit $L = \sum_{i=1}^{K} \lambda_i \bar{X}_i$ où les λ_i sont des nombres fixés. On appellera contraste, cette combinaison linéaire si $\sum_{i=1}^{K} \lambda_i = 0$. Le choix des λ_i nous permettra de comparer les groupes qui nous intéressent. Pour ce faire, on se limite à K - 1 comparaisons qui sont linéairement indépendantes. On peut vérifier l'hypothèse d'égalité des moyennes de certains groupes en comparant la valeur $t = \Sigma \lambda_i \bar{X}_i / S_p$ où S_p^2 est un estimateur de la variance du contraste calculé à partir des groupes concernés. On rejettera l'hypothèse d'égalité si le rapport est supérieur à la valeur critique d'une loi de Student avec Σn_i - K de liberté.

2. RÉSULTATS

On a comme données de base tous les résultats scolaires des étudiants aux niveaux secondaire et collégial ainsi que le classement des étudiants à partie énigmes du test. On cherche à connaître les relations entre le classement et les résultats scolaires en vérifiant s'il existe des différences significatives entre les résultats pour les groupes formés par les étudiants classés concret I, concret II, formel I et formel II. Les analyses de variance qui ont été faites à l'aide du programme " ONEWAY "

de la banque de programmes SPSS [4] ont indiqué qu'il existait des différences significatives entre les moyennes des quatre groupes excepté pour le cours Français 512.

Donc puisque dans la plupart des cours on retrouve des différences significatives entre les moyennes des groupes et que l'on a rejeté l'hypothèse

$$H_o : \mu_{CI} = \mu_{CII} = \mu_{FI} = \mu_{FII}$$

on s'intéresse maintenant à savoir d'où proviennent ces différences. Nous avons donc estimé les contrastes du tableau 1.

Tableau 1: Coefficients des contrastes

Groupes	CI	CII	FI	FII
Contraste 1	1	-1	0	0
Contraste 2	0	0	-1	1
Contraste 3	1	1	-1	-1

Le contraste 1 du tableau 1 permet de tester l'égalité des moyennes des résultats scolaires des étudiants du groupe " concret I " et des étudiants du groupe " concret II ". Par exemple, pour le cours de Physique 422, le tableau 2 nous indique que les moyennes de ces deux groupes sont respectivement 50,4 et 52,7. Le contraste 1 prend donc la valeur 52,7 - 50,4 = 2,3. Le tableau 3 nous indique que cette valeur n'est pas significativement différente de zéro et on conclut donc que les moyennes

des deux groupes pour le cours de Physique 422 sont sensiblement les mêmes. Par contre, les tableaux 2 et 3 nous indiquent que, pour le même cours, la différence entre les groupes " formel I " et " formel II " donnée par le contraste 2 (61,5 - 55,6 = 4,9) est significativement différente de zéro au niveau 1/1 000. Il y a donc une différence importante entre ces deux groupes.

Le troisième contraste permet de vérifier l'hypothèse d'égalité des moyennes entre les groupes d'étudiants classés concrets et ceux classés formels. Bien que le vecteur (1, 1, -1, -1) définissant le contraste 3 soit linéairement indépendant de ceux définissant les contrastes 2 et 3 ((1, -1, 0, 0) et (0, 0, 1, -1)), le test basé sur le contraste 3 n'est pas indépendant de ceux basés sur les contrastes 1 et 2 respectivement étant donné que les tailles des échantillons n'étaient pas égales dans chaque groupe.

L'analyse du tableau 3 nous indique que, dans la plupart des cas, on note une différence importante entre les résultats des étudiants selon le stade qu'il ont atteint à la partie énigmes de l'épreuve E.R.F. On note cependant quelques exceptions (ex. Français 512). Le lecteur intéressé à une analyse plus détaillée de ce tableau pourra consulter le rapport de M. Torkia-Lagacé [3].

Tableau 2 : Moyennes des notes de cours selon le classement à la partie énigme du test

Cours	Concret I	Concret II	Formel I	Formel II
Physique 4?2	50,4	52,7	55,6	61,5
Physique 432	40,5	45,1	49,3	49,1
Physique 522	55,2	59,7	64,4	71,1
Physique 552	51,3	56,5	61,9	68,4
Chimie 442	[illegible],9	62,3	61,6	78,5
Chimie 462	49,3	58,0	64,1	70,0
Chimie 562	46,6	49,5	56,3	64,1
Mathématiques 522	61,4	65,0	69,6	76,5
Mathématiques 532	70,4	72,3	77,1	82,5
Français 512	62,3	66,4	62,8	68,0
Français 522	57,0	60,9	62,9	66,8
Français 532	64,3	67,8	70,5	73,2
* Chimie	- 4,54	- 0,88	3,27	6,61
Biologie	- 0,46	3,05	2,27	4,31
Français	0,60	4,11	4,66	5,92
Mathématiques	- 3,17	- 1,17	4,65	9,49
Physique	- 2,65	- 0,14	2,12	5,76
Philosophie	1,53	4,14	5,03	6,90

* Les notes de cours au collégial ont été calculées comme étant la note de l'étudiant moins la moyenne de son groupe.

Tableau 3 : Relations entre les résultats scolaires des étudiants et leur classement à la partie énigme du test

Notes secondaires	CI-CII	FI-FII	C - F	Nombre d'étudiants
Physique 422		***	***	745
Physique 432	*		***	221
Physique 522	***	***	***	647
Physique 552	***	***	***	1 555
Chimie 442	*	**	***	129
Chimie 462			*	53
Chimie 562	***	***	***	2 780
Mathématiques 522	***	***	***	2 932
Mathématique 532	*	***	***	1 906
Français 512				181
Français 522		***	***	3 125
Français 532		***	***	1 836
Notes collégiales				
Chimie	***	***	***	1 414
Biologie	***		*	943
Français	***	**	***	5 102
Mathématiques	*	***	***	2 913
Physique		***	***	1 016
Philosophie	***	***	***	5 081

Légende

Niveau de signification * 0,05
** 0,01
*** 0,001

BIBLIOGRAPHIE

[1] Cléroux, R. (1976). Analyse de données unidimensionnelles provenant d'une expérience planifiée. Notes de cours IFT 6660, Université de Montréal.

[2] Dixon, W.J. and Massey, Jr. F.J. (1969). Introduction to Statistical Analysis. 3rd Ed., New York, Mc Graw-Hill.

[3] Torkia-Lagacé, M. La pensée formelle des étudiants de collège I; objectif ou réalité ? Collège de Limoilou. 1981

[4] Nie, N.H., Hull, C.H., Jenkins, J.G., Stein Brenner, K. and Bent, D.H. (1975). Statistical Package for the Social Sciences. 2nd Ed., New York, Mc Graw-Hill.